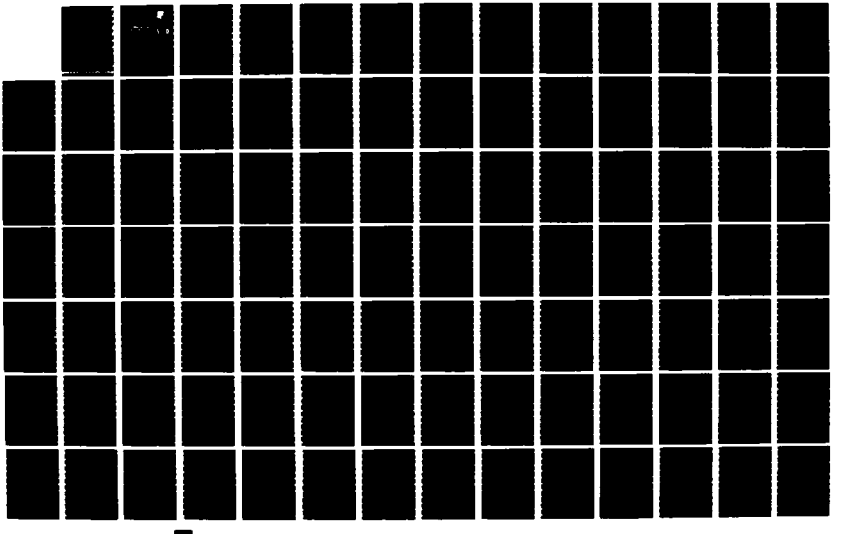
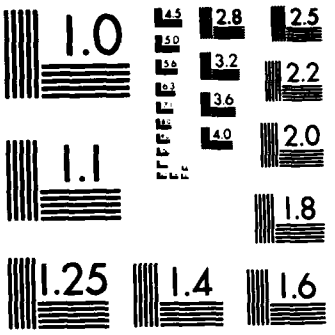
AD-A174 693 NOISE-IMMUNE MULTISENSOR TRANSDUCTION OF SPEECH(U) BBN 1/2
LABS INC CAMBRIDGE MA V R VISWANATHAN ET AL AUG 86
BBN-6114 RADC-TR-86-87 F30602-84-C-0088

UNCLASSIFIED F/G 9/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

RADC-TR-86-87
Final Technical Report
August 1986

NOISE-IMMUNE MULTISENSOR TRANSDUCTION OF SPEECH

BBN Laboratories Incorporated

Vishu R. Viswanathan, Claudia M. Henry, Alan G. Derr,
Salim Roucos, Richard M. Schwartz, and Kenneth N. Stevens

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, NY 13441-5700

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

ADA174693

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS N/A
2a. SECURITY CLASSIFICATION AUTHORITY N/A	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE N/A	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) BBN Report No. 6114	5. MONITORING ORGANIZATION REPORT NUMBER(S) RADC-TR-86-87
6a. NAME OF PERFORMING ORGANIZATION BBN Laboratories Incorporated 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Rome Air Development Center (IRAA)
6c. ADDRESS (City, State, and ZIP Code) 10 Moulton Street Cambridge MA 02238	7b. ADDRESS (City, State, and ZIP Code) Griffiss AFB NY 13441-5700
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Rome Air Development Center 8b. OFFICE SYMBOL (If applicable) IRAA	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F30602-84-C-0088
8c. ADDRESS (City, State, and ZIP Code) Griffiss AFB NY 13441-5700	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 62702F; PROJECT NO: 4594; TASK NO: 15; WORK UNIT ACCESSION NO: 22

11. TITLE (Include Security Classification)
NOISE-IMMUNE MULTISENSOR TRANSDUCTION OF SPEECH

12. PERSONAL AUTHOR(S) Vishu R. Viswanathan, Claudia M. Henry, Alan G. Derr, Salim Roucos, Richard M. Schwartz, and Kenneth N. Stevens

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM May 84 TO Dec 85	August 1986	130

16. SUPPLEMENTARY NOTATION
N/A

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Multisensor Transduction, Speech Recognition Testing in Noise, Noise Reduction, Speech Enhancement (See Reverse)
17	02		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Two types of configurations of multiple sensors were developed, tested and evaluated in speech recognition application for robust performance in high levels of acoustic background noise: One type combines the individual sensor signals to provide a single speech signal input, and the other provides several parallel inputs.

For single-input systems, several configurations of multiple sensors were developed and tested. Results from formal speech intelligibility and quality tests in simulated fighter aircraft cockpit noise show that each of the two-sensor configurations tested outperforms the constituent individual sensors in high noise. Also presented are results comparing the performance of two-sensor configurations and individual sensors in speaker-dependent, isolated-word speech recognition tests performed using a commercial recognizer (Verbex 4000) in simulated fighter aircraft cockpit noise.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION	
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED	
22a. NAME OF RESPONSIBLE INDIVIDUAL Wayne Todd, Capt, USAF	22b. TELEPHONE (Include Area Code) (315) 330-4024	22c. OFFICE SYMBOL RADC (IRAA)

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

For parallel-input systems, selected phonetic discrimination tests involving a feature-based approach were performed to demonstrate the feasibility of using multiple, parallel inputs for high-performance speech recognition. Also, a simple and effective way of using multiple, parallel inputs with an existing speech recognition algorithm was proposed and tested. Results from isolated-word recognition tests show that a two-sensor parallel-input system improves recognition accuracy in high noise substantially over either sensor alone.

ITEM 18. Subject Terms (Continued)

Parallel-Input Speech Recognition Using Multiple Sensors

Table of Contents

List of Figures

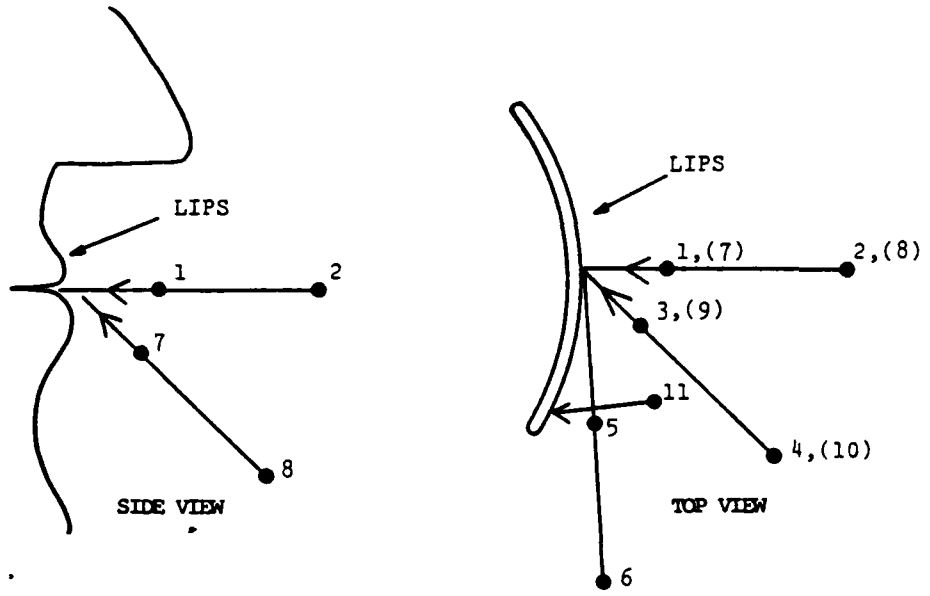
List of Tables

1. INTRODUCTION

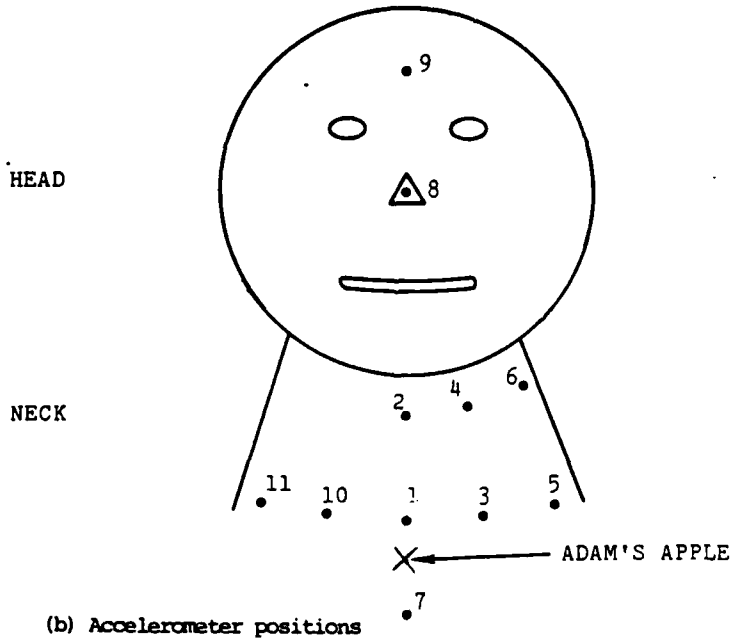
The overall objective of this research was to develop, test, and evaluate in a speech recognition application multisensor configurations of existing sensors for transducing speech that is more immune to acoustic background noise than is possible with any single microphone. In this section, we review briefly the results of the previous RADC-sponsored multisensor speech input project that provided the basis for the present effort, state the specific goals of this research, present the highlights of the work, and describe the organization of the report.

1.1 Review of Multisensor Speech Input Project

In the previous RADC-sponsored multisensor research [1, 2], we performed detailed measurements of the sound field in the vicinity of the mouth and neck during speech using pressure and pressure gradient (noise-cancelling) microphones and an accelerometer that measures the skin vibrations. We used the first-order pressure gradient output (denoted in this report as M12) from a specially-constructed array of three closely-spaced electret microphones, ElectroVoice's first-order gradient microphone EV 985, Vought's prototype second-order gradient microphone, the electret pressure microphone M1 of our three-microphone array, and a reference Bruel and Kjaer condenser pressure microphone (located about one foot from the speaker's lips). The accelerometer we used is Model 501, which is marketed by Vibro-Meter Corporation (formerly BBN Instruments Corporation). We performed the sound-field measurements in a noise-free anechoic chamber, using five talkers, a specially selected set of speech materials, and eleven selected positions for each sensor, which are illustrated in Fig. 1. Compared to the microphones we used, the accelerometer is essentially insensitive to acoustic noise, especially at low frequencies. We investigated in detail the data from one male talker both using long-term and short-term spectral analyses and also using, as an objective measure of speech intelligibility, the articulation index, which we computed assuming ambient noise typical in the cockpit of an F-15 fighter aircraft. From the results of this investigation, we determined the best position for each sensor; also, we



(a) Microphone positions



(b) Accelerometer positions

FIG. 1. Sensor positions used in our sound-field measurements during speech.

developed a two-sensor configuration involving the accelerometer attached to the speaker's throat and one of the gradient microphones located in front of the lips. The two-sensor signal is the sum of filtered and amplitude-adjusted versions of the two individual sensor signals, with the accelerometer providing the low-frequency information and the gradient microphone providing the high-frequency information. Results from formal speech intelligibility and quality tests in simulated F-15 aircraft cockpit noise showed clearly that each of the two-sensor signals under test outperformed the signal from the gradient microphone alone and that the performance improvement generally increased with the noise level.

1.2 Research Goals of this Project

The multisensor configuration that we have sought to develop must provide additional, new acoustic information not presently used, a more reliable and accurate transduction of the presently used information, or a more robust way of extracting the same information in a hostile acoustic environment. Our goal in this project was to develop, test, and evaluate in a speech recognition application two types of multisensor configurations. The first type is called the single-input multisensor configuration, which combines the individual sensor signals to provide a single speech signal input to any speech processing system. (Notice that the two-sensor systems developed in the previous multisensor project belong to this first type.) The second type is called the parallel-input multisensor configuration, which provides several parallel inputs; these input signals are analyzed to extract features or parameters for use in applications such as speech recognition and speech-training aids [3]. The general objective of this project was to achieve noise-immune speech transduction with single-input multisensor systems and noise-immune feature extraction with parallel-input multisensor systems. Specific objectives of our work on single-input multisensor systems were to develop and test a number of multisensor systems, select the most promising ones, and evaluate the selected systems in simulated F-15 aircraft cockpit noise, using a commercial speech recognition device. The specific objective of our work on parallel-input multisensor systems was to demonstrate the feasibility of a speech recognition system that effectively uses the multiple, parallel inputs.

1.3 Highlights of the Work

Given below is a list of the highlights of our work:

- Long-term and short-term spectral analyses and articulation index study of the previously measured data of one male and one female speaker, transduced using different sensors at different near positions. The results of this investigation were used 1) to determine over the two speakers the extent of variability of the best location and the spectral properties of each sensor and 2) as a basis for developing new multisensor configurations (Section 2).

- Development and testing of a method of spectrally shaping the accelerometer signal by emphasizing high-frequency amplitudes relative to low-frequency amplitudes, in an attempt to improve the performance of the single-input two-sensor system involving an accelerometer and a gradient microphone (Section 3).

- Development of a new single-input two-microphone system, with the Vought second-order gradient microphone providing the low-frequency information and the first-order gradient microphone M12 providing the high-frequency information. The two-microphone signal has a higher overall signal-to-noise ratio than either of the gradient microphones (Section 4).

- Development of several additional single-input multisensor systems involving all or different subsets of the sensors, throat accelerometer, Vought, M12, and EV 985 located under one nostril (Section 5).

- Multichannel tape recording of the following items from one male and one female talker, in the quiet and in 95 dB and 115 dB levels of simulated F-15 aircraft cockpit noise: Diagnostic Rhyme Test (DRT) word lists for speech intelligibility testing, a set of six sentences for speech quality testing, and a 20-word Texas Instruments (TI) vocabulary and a specially selected 44-word minimal pairs vocabulary for isolated-word speech recognition testing (Section 6).

- Formal speech intelligibility and quality testing of selected single-input two-sensor systems and individual sensors in 95 dB and 115 dB noise. The two-sensor systems tested produce essentially the same DRT scores and quality ratings in 95 dB and

much higher DRT scores and quality ratings in 115 dB, as compared to the constituent individual microphones (Section 6).

- Detailed speech recognition performance evaluation of selected single-input two-sensor systems and individual sensors in noise, using the Verbex 4000 recognizer in the isolated-word mode. For this evaluation, we used the TI vocabulary and two subsets of our minimal pairs vocabulary. Because of the training problems of the Verbex unit in high noise and because of the limitations of the vocabularies we used, we cannot draw strong definitive conclusions from the results of this evaluation. However, the results provide us with sufficient evidence to recommend the use of a two-sensor system involving the accelerometer and a gradient microphone, together with a variable cutoff lowpass filter to increase the extent of band-limiting of the two-sensor signal as a function of the ambient noise level (Section 7).

- Feasibility demonstration of parallel-input multisensor speech recognition, using selected phonetic discrimination tests. We used BBN's versatile Acoustic-Phonetic Experiment Facility for this demonstration and obtained substantially higher phonetic discrimination accuracy for a feature-based parallel-input multisensor system we investigated than for the gradient microphones we used (Section 8).

- Investigation of a long-vector approach to parallel-input multisensor speech recognition. In this approach, we formed, on a frame-by-frame basis, a composite (or long) vector of parameters by simply collecting together the parameters extracted from each of the parallel inputs and evaluated the long-vector data using BBN's research speech recognition system, which employs vector quantization and a discrete hidden Markov model. The results of this investigation show a substantially higher recognition accuracy for a parallel-input system consisting of M12 and a throat accelerometer than for either constituent sensor (Section 9).

1.4 Organization of the Report

Section 2 presents the results of analyses we performed on previously measured data for two speakers. Sections 3-7 deal with single-input multisensor systems, and Sections 8 and 9

deal with parallel-input multisensor systems. It is important to keep in mind the two groupings of sections since we frequently do not use the term 'single-input' or 'parallel-input' when we refer to a multisensor system, although which type of system we mean should be clear from context. Contents of Sections 3-9 were highlighted in the previous subsection. In Section 10, we provide a summary and present major conclusions. Appendix A is a listing of the contents of the quality-test database, TI vocabulary, and minimal pairs vocabulary. Appendix B describes the acoustic-phonetic features we used in phonetic discrimination tests presented in Section 8.

2. ANALYSIS OF TWO-SPEAKER DATA

As we noted in Section 1.1, speech data collected in a noise-free anechoic chamber as part of the previous multisensor speech input project was digitized and analyzed for one speaker (male, KK) only [1]. In this project, we digitized the data for a second speaker (female, BF), and investigated the data of both speakers using long-term and short-term analyses. The objectives of this investigation were to 1) examine if the spectral properties we reported in [1] for one speaker continue to be valid for the second speaker; 2) determine the extent to which location of each sensor needs to be tailored to the individual speaker; and 3) gather supportive data for developing both single-input and parallel-input multisensor systems.

2.1 Informal Listening Tests

For speaker BF, we played out the digitized waveform files for all eleven accelerometer positions to determine, through informal listening, which positions yielded the most intelligible speech. All the files were highpass filtered at 200 Hz on playback to reduce "boominess" [1]. We found that position 3 was the most intelligible, closely followed by positions 11, 5, and 10. As was the case with speaker KK, position 7 was barely intelligible. (The results for informal listening tests conducted with KK's data are discussed in [1], pages 50-51.) The accelerometer signal in position 10 was corrupted by scratching noises, possibly caused by improper mounting of the accelerometer or by its coming into contact with a shirt collar; also, some of the vowel sounds recorded in position 10 had more of a buzzing quality than those recorded in positions 3, 11, and 5. Once again, this could have been caused by improper mounting of the accelerometer. Of the 4 top-rated positions mentioned above, position 3 produced the least "boominess". We expect that with proper mounting of the accelerometer, positions 3 and 10 should be equivalent. Thus, the position that produced the most intelligible speech was the same (3 or 10) for speakers KK and BF.

2.2 Articulation Index Analysis

We computed the long-term spectra for the two speakers over five sentences, for all the near positions (1, 3, 5, 7, 9, and 11) and for each of the four microphones, M1, M12, EV 985, and Vought. We examined the long-term speech spectra and the long-term F-15 aircraft cockpit noise spectra for the four microphones; the noise spectra were computed from the noise-only responses of the microphones, which were measured in the previous project [1]. The results of this investigation provided the basis for developing a two-microphone configuration, which is described in detail in Section 4.

We then performed articulation index analysis using the procedure described in [1]. We recall that the articulation index (AI) is an objective measure of the intelligibility of speech in noise. As noise, we considered the ambient noise typical in the F-15 fighter aircraft cockpit (see Fig. 18 in [1]). We computed the AI scores for speakers KK and BF at each of the 3 noise levels, 85 dB, 107 dB, and 114 dB. For each speaker, we evaluated the 85 dB case twice, once with normal voice and once with raised voice (6 dB increase in the speech level); for the other two noise levels, we used raised voice only.

The AI scores are given in Table 1 for speaker KK and in Table 2 for speaker BF. Since positions 5 and 11 produced substantially lower AI scores than did the other four near positions, we excluded these two positions in Tables 1 and 2. Notice that we have rank-ordered the four positions in terms of their AI scores, for each microphone.

We make several observations. First, let us consider the issue of the best position for each microphone. For speaker KK, we find from Table 1 that the rank-ordering of the four positions is the same for all four cases (a)-(d), for each of the first three microphones. For Vought, the two lower ranked positions 3 and 9 are interchanged for (c) and (d) relative to (a) and (b). Also, position 7 is the best for all four microphones and for all four cases (a)-(d). The

Rank	M1	M12	EV 985	Vought
	Pos. Score	Pos. Score	Pos. Score	Pos. Score
1	7 0.530	7 0.813	7 0.764	7 0.781
2	9 0.496	3 0.806	3 0.749	1 0.766
3	3 0.486	9 0.782	9 0.738	9 0.698
4	1 0.477	1 0.727	1 0.729	3 0.692
(a) 85 dB noise level and normal voice				
1	7 0.720	7 0.900	7 0.891	7 0.866
2	9 0.686	3 0.895	3 0.878	1 0.852
3	3 0.676	9 0.879	9 0.869	9 0.808
4	1 0.667	1 0.859	1 0.867	3 0.800
(b) 85 dB noise level and raised voice				
1	7 0.058	7 0.363	7 0.271	7 0.437
2	9 0.043	3 0.348	3 0.258	1 0.407
3	3 0.030	9 0.329	9 0.246	3 0.339
4	1 0.027	1 0.260	1 0.239	9 0.316
(c) 107 dB noise level and raised voice				
1	7 0.000	7 0.172	7 0.089	7 0.290
2	9 0.000	3 0.159	3 0.083	1 0.266
3	3 0.000	9 0.153	9 0.080	3 0.204
4	1 0.000 (cannot rank)	1 0.088	1 0.072	9 0.135
(d) 114 dB noise level and raised voice				

Table 1. AI scores for the four best positions of the different sensors, for speaker KK.

Rank	M12		EV 985		Vought	
	Pos.	Score	Pos.	Score	Pos.	Score
1	7	0.612	9	0.534	7	0.570
2	3	0.600	7	0.516	1	0.561
3	9	0.587	1	0.490	3	0.539
4	1	0.540	3	0.474	9	0.517
(a) 85 dB noise level and normal voice						
1	7	0.758	9	0.708	7	0.694
2	3	0.746	7	0.698	1	0.686
3	9	0.741	1	0.667	3	0.658
4	1	0.702	3	0.654	9	0.648
(b) 85 dB noise level and raised voice						
1	7	0.160	9	0.104	7	0.252
2	3	0.151	7	0.080	1	0.238
3	9	0.147	1	0.072	3	0.228
4	1	0.112	3	0.069	9	0.202
(c) 107 dB noise level and raised voice						
1	3	0.064	9	0.021	7	0.151
2	7	0.062	1	0.016	1	0.143
3	9	0.048	3	0.014	3	0.135
4	1	0.039	7	0.012	9	0.118
(d) 114 dB noise level and raised voice						

Table 2. AI scores for the four best positions of the different sensors, for speaker BF.

second best position is 9 for M1, 3 for M12 and EV 985, and 1 for Vought, again uniformly for all four cases. Considering speaker BF, we first see from Table 2 that the AI scores are consistently lower than those given in Table 1 for speaker KK. (The AI scores for M1 are not given in Table 2 as they were quite low, especially for cases (c) and (d).) Nonetheless, we see from Table 2 that the rank ordering of the positions is uniform over all four cases for Vought, and for cases (a)-(c) for M12 and EV 985. The best position overall is 7 for M12 and Vought and 9 for EV 985. The second best position overall is 3 for M12 and 1 for Vought. For EV 985, the AI scores are extremely low for case (d). For cases (a), (b), and (c), the second best position for EV 985 was 7, and it was only slightly worse than the best position 9. To examine why position 3 of EV 985 seems to be inferior for BF unlike for KK, we have given in Table 3 the actual microphone distances we used in the original anechoic chamber measurements. We find that the difference in EV 985 distance between positions 3 and 9 was 0.5 cm for BF and was only 0.1 cm for KK, with position 3 being farther from the mouth. Notice also that positions 7 and 9 were equidistant from the mouth for KK, but position 9 was closer to the mouth for BF. This may explain the discrepancy we mentioned above for EV 985. In general, position 7 appears to be the best position for all microphones, with the second choice being either position 3 or position 9 for M12 and EV 985, and position 1 for Vought. (We raise caution here by noting that positions 1 and 7, which are directly in front of the lips, are likely to be very sensitive to puffnoise or breath noise that occurs for plosive sounds. See Section 6.1 below.)

Second, we consider the sensitivity of the microphones to orientation. We find from Tables 1 and 2 that in general, positions 3 and 9 produce very similar AI scores for all microphones, and that positions 1 and 7 produce similar AI scores only for Vought. Of course, the differences in microphone distances could have been partly responsible for this result. Since we always located the microphones as close to the mouth as possible (without the puffscreen touching the lips for sounds such as [u] in boot), we conclude that the Vought microphone, located either directly in front of the talker (position 1 or 7) or to the side at a 45-degree angle (position 3 or 9), is not sensitive to slight errors in orientation. On the other hand, this is true for M12 and EV 985 only when they are located at the 45-degree angle.

Position	BBN Array (M1 & M12)		EV 985		Vought	
	KK	BF	KK	BF	KK	BF
1	0.7	0.7	2.2	2.4	2.4	2.5
3	0.6	0.7	2.4	2.7	2.2	2.9
7	0.5	0.7	2.3	2.4	2.0	2.3
9	0.6	0.8	2.3	2.2	2.5	2.9

Table 3. Distances (in cm) of the three microphones to the center of the mouth for four near positions, for speakers KK and BF.

Third, we recall the rule of thumb that states that an AI score above 0.4 leads to intelligible speech. Considering speaker KK, even the best position produces AI scores below 0.4 for the high noise level of 114 dB; for 107 dB, only Vought in position 7 produces an AI score above 0.4. (These results, of course, provide one motivation for developing multisensor configurations.)

Fourth, we observe that M12 produces consistently higher AI scores than Vought does for the low noise level of 85 dB. The nearer location of the microphone array (see Table 3) is responsible for this result. At higher noise levels, the better noise-cancelling property of the Vought microphone helps to improve its AI score relative to that of M12.

2.3 Short-Term Spectral Analysis

In this study, we attempted to identify the strengths and weaknesses of each sensor in each position by observing its response to individual phonemes rather than by observing its long-term average speech response. In this way, we hoped to develop sensor configurations

and methods of mixing the signals together that would improve the noise immunity for a wide variety of speech sounds and for different noise conditions. By observing certain gross characteristics of the output signals from various sensors, we sought to distinguish among different classes of sounds (vowels, fricatives, nasals, etc.). With this information, we believed that it might be possible to use different signal mixing rules for different classes of sounds in order to exploit the strengths of the various sensors. We recall that the sensor data used in the short-term spectral analysis had been collected in a noise-free anechoic chamber.

2.3.1 Gradient Microphones

To determine the strengths and weaknesses of the M12 and Vought microphones in different positions, we analyzed the spectral responses of these microphones to individual phonemes for speaker KK. Specifically, we concentrated on the near positions 1, 3, 7, and 9 for the following phoneme categories: vowels, voice bars, nasals, unvoiced fricatives, voiced fricatives, and plosives. We also studied positions 5 and 11 for unvoiced fricatives for M12 only. For each microphone and position, we used two criteria to evaluate each short-term spectrum associated with a particular phoneme. First, we studied the extent to which the short-term spectra of the microphone and the reference microphone resembled each other; in general, the greater this resemblance, the more accurate the microphone's response to the phoneme in question. Second, we examined the signal-to-noise ratio performance by comparing the microphone speech spectrum's level with the long-term cockpit noise spectrum's level, concentrating on overall noise levels of 107 and 85 dB. From this investigation, we made the following observations:

1. For a significant number of phonemes, position 1 for the Vought seemed to be marginally better than positions 3, 7, and 9. For M12, on the other hand, position 1 tended to be slightly worse than positions 3, 7, and 9. In general, however, for each microphone and phoneme, differences among the various positions were not usually dramatic, and few indications of any one position's superiority for a particular phoneme became evident for any of the phonemes we studied. Since there seemed to be no strong phoneme dependence apparent in the assessment of

one microphone position relative to the others, we ruled out the possibility of performing a variable, phoneme-dependent mix of the microphone signals in a single-input multisensor system.

2. Vowels were fairly strong relative to the noise, while voice bars and nasals were very weak. Fricatives and plosives fell between these two extremes. Because the sound source for voice bars (radiation from throat surface) and nasals (radiation from nostrils) is not in the near field of a gradient microphone, the amplitudes of the transduced signal for these sounds are expected to be substantially smaller (in relation to an adjacent vowel) for M12 or Vought than for the reference microphone. We found this result to be true in our investigation (see also Section 4.6 in [1]). For improved transduction of voice bars and nasals, we therefore suggest the use of a gradient microphone with a second sensor, e.g., an accelerometer (see the next subsection for more discussion).

3. M12's speech response was usually stronger relative to its noise response at higher frequencies (above 2 kHz) than Vought's, while the reverse was true at low frequencies. This observation is consistent with the results of our long-term spectral analysis (see Section 4).

4. Because more air tends to be expelled from the sides of the mouth during the production of fricatives, we decided to investigate M12's "side positions" 5 and 11 for fricatives to determine if these positions were more useful for fricatives than the other near positions. In our comparisons, position 5 appeared to be better than positions 1 and 11 but worse than positions 3, 7, and 9. Therefore, no advantage for fricatives was gained by transducing them with a microphone in a "side position."

We also analyzed speaker BF's short-term spectra for Vought and M12 for several phonemes. We discovered that the differences between positions for a given microphone and phoneme were even less pronounced than they were for KK; position 1 for Vought showed no noticeable superiority over the other positions, and position 1 for M12 was not noticeably inferior to the other positions. Therefore, we chose to base our decisions concerning the "best" position for each microphone on the articulation index scores (see Section 2.2) and the results of informal listening tests (Section 6.1).

2.3.2 Accelerometer

In our study of short-term spectra for the accelerometer, our goal was to determine how well different phonemes are transduced by the accelerometer attached to the throat in position 10, as well as to search for features that could be useful in a feature-based parallel-input speech recognition system; recall from Section 2.1 that we had already chosen position 10 (or its symmetric position 3) as the "best" position for the throat accelerometer through informal listening tests. In addition, we briefly investigated the feasibility of using the nasal accelerometer signal (position 8) for the transduction of nasal sounds; this investigation was particularly important because, as mentioned above, the gradient microphones do not transduce nasals well.

Because we expected the throat accelerometer to provide no useful information for the unvoiced phonemes, we concentrated our attention on the voiced phoneme classes: voiced fricatives, vowels, voiced plosives (voice bars), and nasals. To determine how well the accelerometer transduces these phonemes, we compared its short-term spectra with the short-term spectra for the same phonemes derived from a reference microphone placed one foot from the talker. For a given phoneme, the more closely the spectra matched, the better we assumed to be the accelerometer's ability to transduce that particular phoneme. We note that all data used in this study was recorded in the quiet to facilitate comparisons with the reference microphone. We expected that the accelerometer's relative insensitivity to acoustic noise would keep its spectral shape fairly constant, regardless of the level of ambient noise.

Because of the sharp lowpass roll-off inherent in the accelerometer signal, we assumed that the kind of spectral shaping described in Section 3, which emphasizes the high frequencies relative to the low frequencies, was applied to each accelerometer spectrum so that its overall shape did not differ too much from the shape of the reference microphone spectrum for the same phoneme. A discussion of our findings in this study follows.

The spectrum of a typical voiced fricative [z] measured at the reference microphone, plotted in Fig. 2, shows regularly spaced harmonic peaks for frequencies below approximately

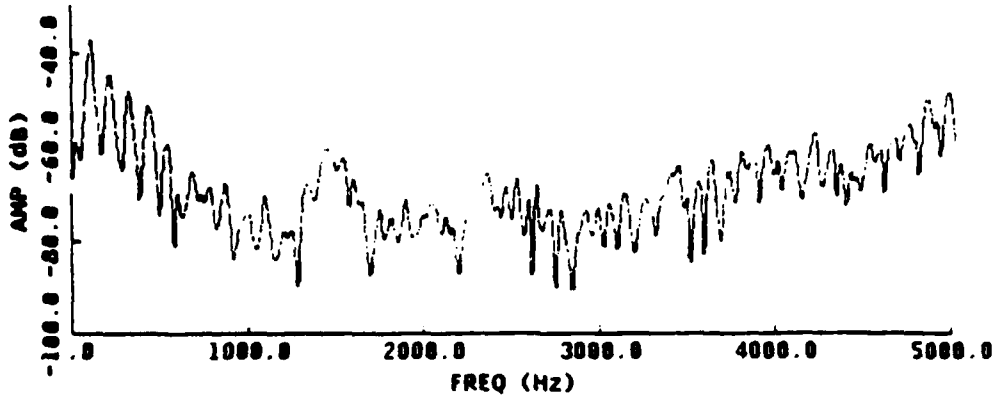


FIG. 2. Spectrum of [z] transduced by the reference microphone at 1 foot.

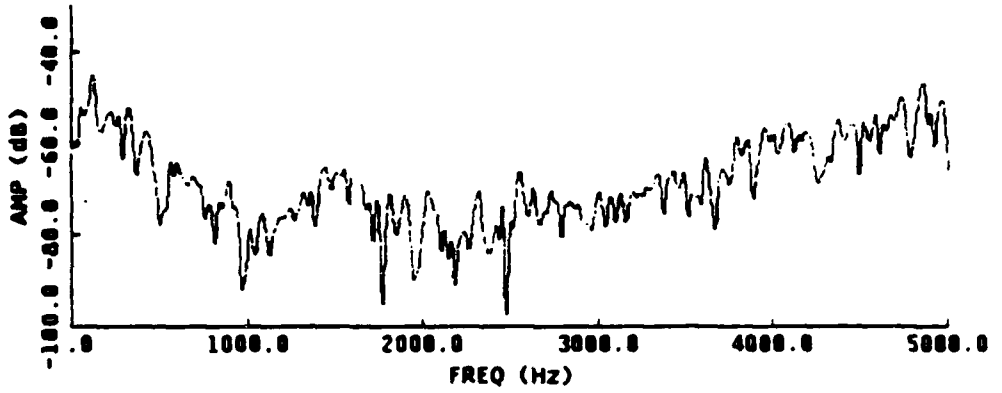


FIG. 3. Spectrum of [z] transduced by M12 in position 7.

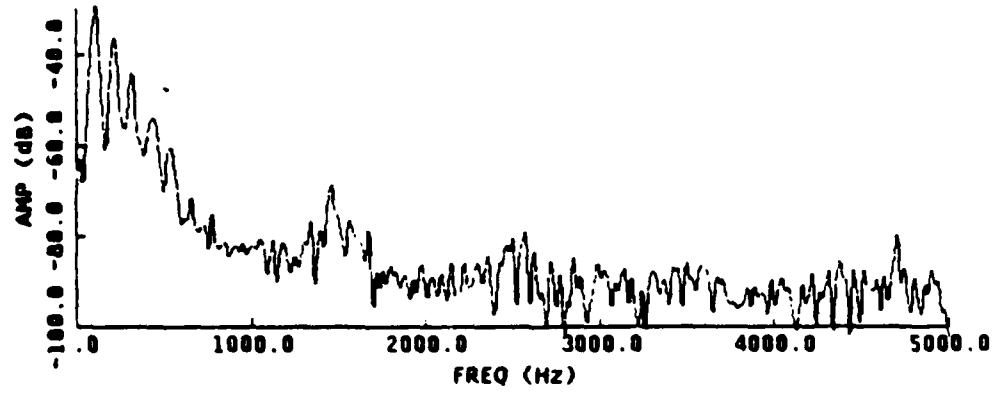


FIG. 4. Spectrum of [z] transduced by the accelerometer in position 10.

700 Hz followed by a gradual upward slope with irregularly spaced peaks at higher frequencies. The low-frequency part of the spectrum indicates the presence of voicing due to periodic vibration of the vocal cords. The high-frequency part of the spectrum is characteristic of a shaped noise source, in this case turbulence noise created by forcing air through a constriction in the vocal tract. The spectrum of this same phoneme as measured by the BBN first-order gradient microphone M12 in position 7, plotted in Fig. 3; shows that the low-frequency part does not exhibit periodic peaks, but the high-frequency part closely follows the shape of the reference microphone spectrum. However, the spectrum of this sound measured using an accelerometer in position 10, plotted in Fig. 4, shows the regular low-frequency peaks very clearly; the high-frequency part of the spectrum is nearly flat. These results can be explained as follows. Because the constriction in the vocal tract allows very little air past it, most of the vibrating air (due to voicing) is trapped behind the constriction. Therefore, the turbulence noise is the major component of the signal that passes the lips. The voiced part of the sound that a listener hears results from the vocal cords' vibration being transmitted through the skin of the throat and then radiated into the air. The accelerometer, as expected, transduced the voicing information well but was not as sensitive to the frication part. M12, on the other hand, transduced the frication well, but was not as sensitive to the voicing part, as the source in this case is not in the near field of M12. The reference pressure microphone, as expected, transduced both the voiced and the fricated parts well. These results also indicate that the single-input two-sensor configuration consisting of the accelerometer and M12 will transduce the voiced fricatives well, since the spectrum of the two-sensor signal will contain the low-frequency periodic peaks of the accelerometer signal and the high-frequency spectral slope of the M12 signal.

During informal listening tests of the accelerometer signal (see Section 2.1), we observed problems with transducing certain vowel sounds when the accelerometer was used in position 10, which is generally the most intelligible position. When listening to the accelerometer output from this position, the phoneme [u] (as in "bOOt") was consistently confused with [i] (as in "bEAt"). It is easy to see from plots of the smoothed spectra of the accelerometer signal and the reference microphone signal why this confusion occurred. (We used the linear

prediction or LPC method for smoothing the spectra.) For the phoneme [i], Fig. 5 shows that the reference spectrum has two formant peaks below 2500 Hz, one at approximately 300 Hz and one near 2000 Hz. The accelerometer spectrum for this sound shows the same peaks. For the phoneme [u], Fig. 6 shows that the reference spectrum has three peaks below 2500 Hz, approximately at the frequencies 300 Hz, 1000 Hz, and 2200 Hz. However, in the accelerometer spectrum, the formant peak near 1000 Hz has dropped out, leaving a spectrum that looks very similar to the spectrum of [i]. This explains the [u] to [i] confusion mentioned above. We believe that the loss of the second formant in the accelerometer signal for [u] was caused by the following mechanism. The formant peaks in the speech spectrum are the result of resonances within cavities formed by the articulators (tongue, lips, teeth, and roof of the mouth) along the vocal tract. The second formant in the phoneme [u] is the result of a resonance in the front cavity formed between the tongue and the lips. In the phoneme [i], this cavity is opened up at the lip end, and hence this resonance does not occur. Normally, when we listen to speech, we hear it as it radiates from the lips. By attaching the accelerometer to the side of the throat, we can effectively listen to the speech as it would sound at the vocal cord end of the vocal tract. Because of the constriction formed by placing the middle of the tongue near the roof of the mouth for these sounds, the front cavity resonances are decoupled from the rear cavity, and thus front cavity resonances are absent or severely attenuated in the accelerometer signal. This effect can also be observed for other phonemes for which the front and back cavities are decoupled or loosely coupled. To determine if, for any other accelerometer positions, the vowel formants are picked up more clearly, we briefly investigated the short-term spectra for vowels transduced in positions other than 10. We found, however, that the other accelerometer positions were no better than position 10 for vowels.

We note that the missing information in the accelerometer signal is unretrievable. This underscores the importance of using the accelerometer signal together with another sensor signal (e.g., gradient microphone signal).

For voice bars, we found that the spectrum of an accelerometer in position 10 (or position 3) matched the spectrum of the reference microphone signal quite well at low frequencies.

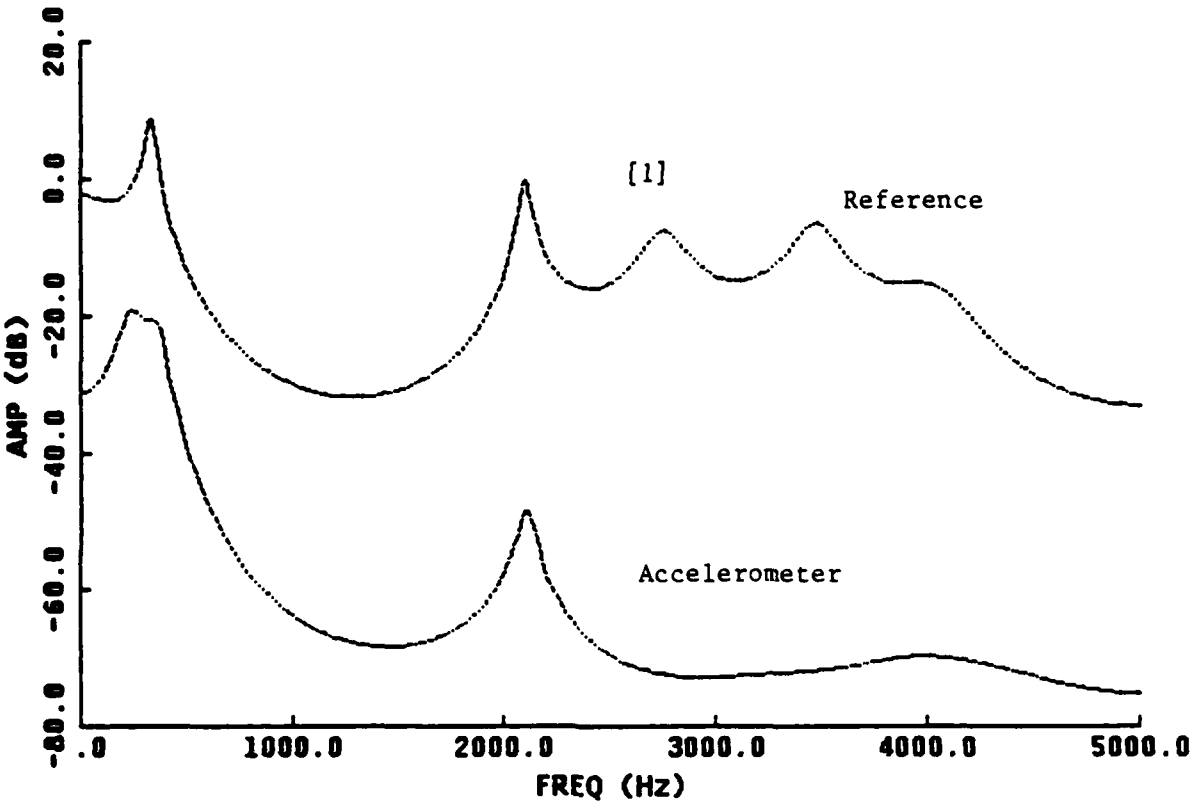


FIG. 5. Comparison of LPC-smoothed spectra for reference microphone and accelerometer for vowel [i] in "beat".

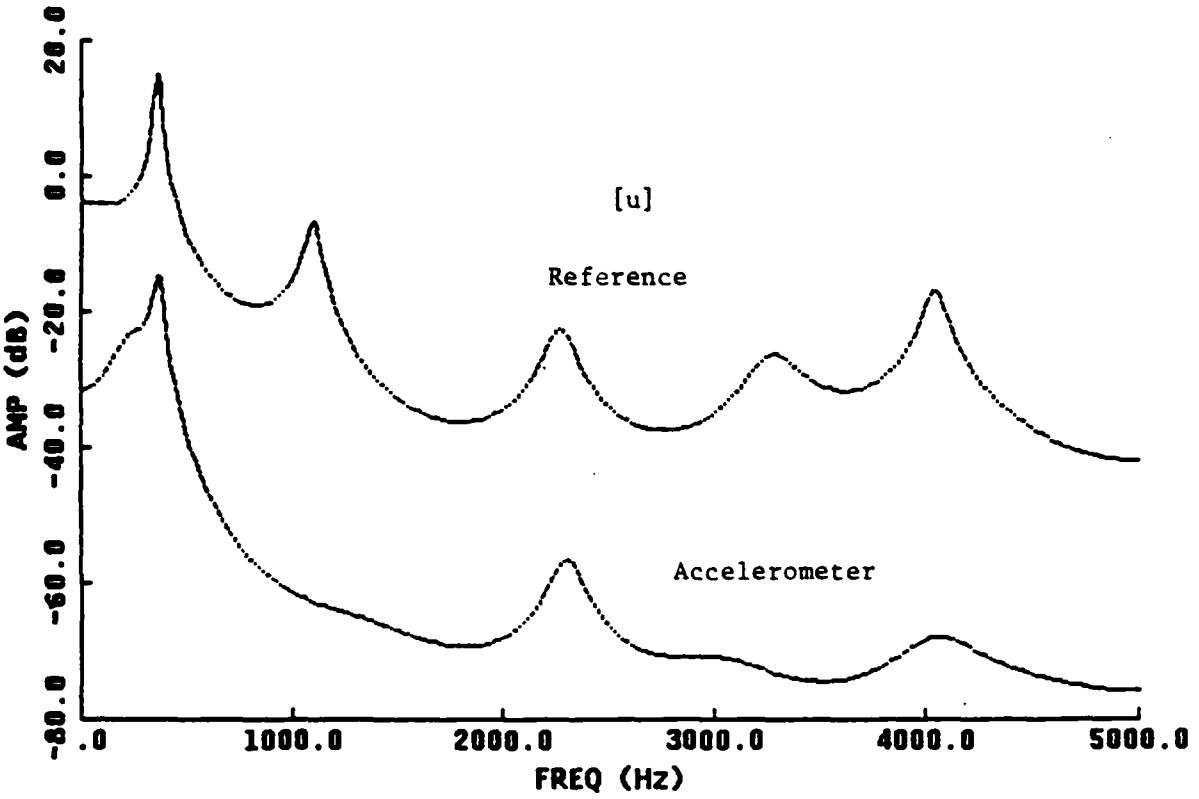


FIG. 6. Comparison of LPC-smoothed spectra for reference microphone and accelerometer for the vowel [u] in "boot"

Since most of the energy for voice bars is in frequencies lower than 1 kHz, this suggests that the single-input two-sensor system involving the accelerometer and a gradient microphone will transduce the voice bars well.

In the process of analyzing the short-term spectra for speaker KK, we observed several spectral features that appeared promising for use in the parallel-input multisensor system. Most notably, a sharp resonant peak was observed in the LPC-smoothed spectra of the throat accelerometer signal for voice bars, voiced fricatives, and nasals, roughly in the range of 1 to 2 kHz; the frequency location of this peak seemed to be phoneme-dependent. We then studied the short-term LPC-smoothed spectra of the same sounds for female speaker BF, for accelerometer positions 3 and 10. Some of the spectra, such as for the voiced fricatives [zh] and [v], yielded a fairly strong peak; but for some other phonemes, particularly voice bars, the peak was either far less pronounced or nonexistent. In other words, for speaker BF this peak demonstrated far less consistent behavior than it did for KK.

The use of the accelerometer signal for accurate pitch and voicing extraction has been demonstrated recently as part of another government-sponsored contract at BBN [4]. Position 7 (just below the glottis) was used in that project. Accurate pitch and only 1% voicing error were reported for a database of 50 sentences from 3 males and 3 females. We performed some limited testing involving only a few sentences of speech transduced by the accelerometer in position 10 and observed, through visual examination of speech waveforms and extracted pitch data, approximately the same 1% voicing error and accurate pitch.

For the nasal accelerometer (position 8), we found that the spectrum of its output signal closely matched the spectrum of the reference microphone for nasals and nasalized sounds except for the high-frequency roll-off characteristic of the accelerometer signal. For other sounds, the output signal level of this accelerometer was very low. Because the nasal accelerometer transduces nasals well, we concluded that including it in a multisensor system would probably help in the human and/or machine recognition of nasal sounds.

We studied briefly the feasibility of a nasality detector based on the frame energy of the nasal accelerometer signal. We developed a simple algorithm that compares the frame energy

with an adaptive threshold; if the threshold is exceeded, the frame in question is declared nasal or nasalized. We conducted only limited tests, and our nasality detection algorithm seemed to perform quite well, indicating the feasibility of this approach.

In conclusion, we found that the throat accelerometer signal would be very helpful in detecting the presence of voiced sounds. Also, it showed promise for actually discriminating among some voiced sounds, although it demonstrated some weaknesses as well, as shown by the drop-out of the second formant peak in [u]. Short-term spectral analysis of the nasal accelerometer signal indicated that it is useful for the transduction of nasal sounds. As an important reminder, we note that the properties of the accelerometer signal will continue to hold even in high noise since the accelerometer is essentially insensitive to acoustic noise.

3. SPECTRAL SHAPING OF ACCELEROMETER SIGNAL

As noted in [1], the accelerometer signal is severely attenuated at high frequencies because of skin conduction losses. However, results from informal listening tests on the accelerometer signal processed through a variable-cutoff highpass filter showed that it contains useful information (intelligible speech), though low in level, even at frequencies above 3 kHz. Spectral analysis of the accelerometer signal also indicated the existence of high-frequency signal amplitudes larger than the accelerometer's noise floor, for voiced sounds. Encouraged by these results, we investigated a method of spectrally shaping the accelerometer signal; this method emphasizes the high-frequency amplitudes relative to the low-frequency amplitudes, to enhance the low-level high-frequency information and to reduce the unnatural quality (e.g., boominess) due to the exaggerated low-frequency amplitudes. The spectrally shaped accelerometer signal would then be combined with a gradient microphone's signal, in an attempt to improve the single-input two-sensor system developed in the previous project [1] and mentioned above in Section 1.1.

We determined for two speakers, one male (AD) and one female (CH), the spectral shaping function through informal listening tests on the accelerometer signal processed using an adjustable one-third octave filterbank. The accelerometer was placed in position 10 for both speakers (refer to Fig. 1). The spectra of the resulting shaping functions are shown in Fig. 7. As shown in the figure, the shaping we chose for either speaker provides a sharp highpass at about 800 Hz, a 5 dB/octave boost over 800-2800 Hz, a flat response over 2.8-4.7 kHz, and a sharp lowpass at 4.7 kHz. The only appreciable difference in the spectral shapes for the two speakers occurred at frequencies below 1 kHz, where CH's signal was attenuated more than AD's. The shaped version of speaker KK's accelerometer signal recorded in 107 dB SPL cockpit noise contained a fair amount of high-frequency noise; the noise level was reduced appreciably when we lowpass filtered the shaped signal at about 3.2 kHz. In subsequent tests, we used the shaper-lowpass filter cascade. Speaker AD's long-term spectra of speech transduced by the accelerometer in position 10 both with and without the "optimal" shaping are shown in Fig. 8.

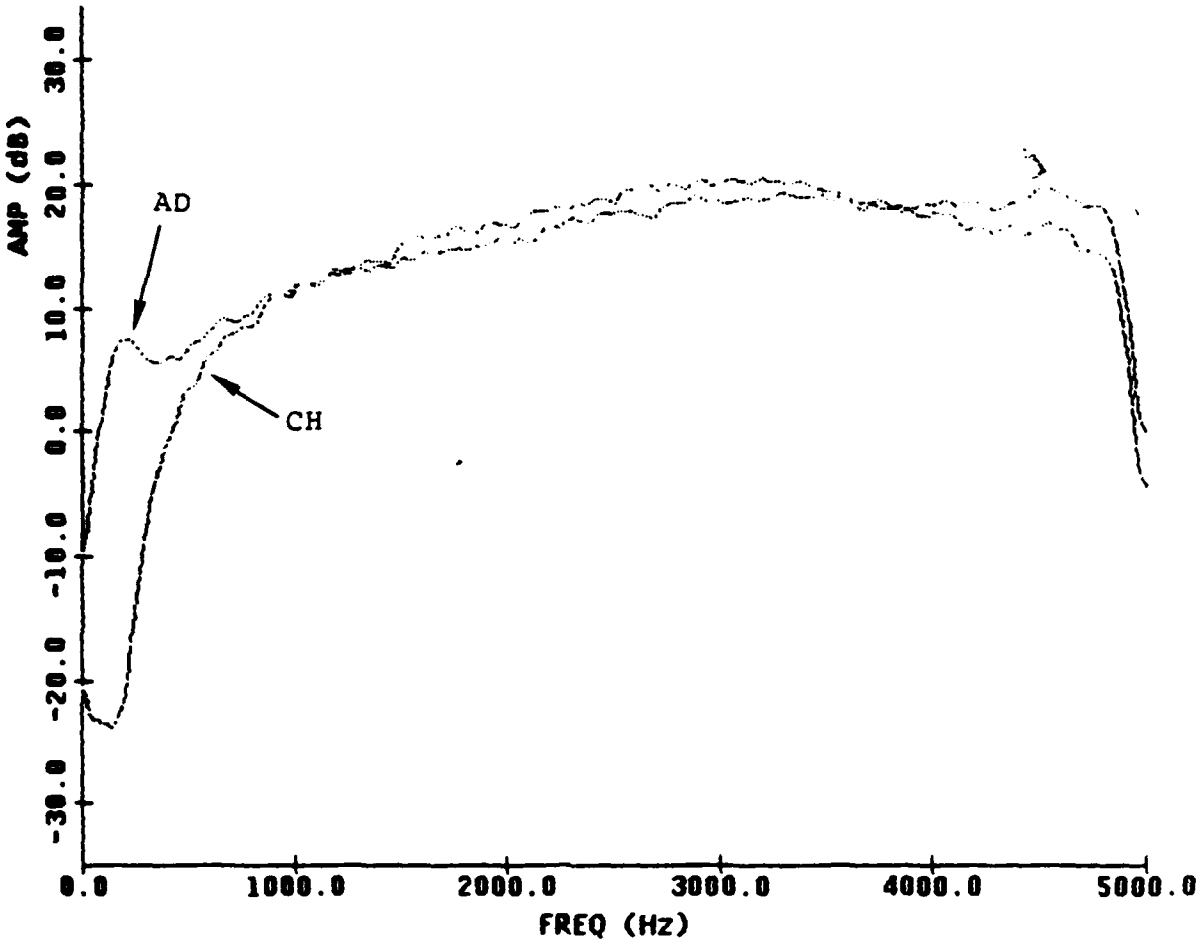


FIG. 7. "Optimal" accelerometer filter shapes for speakers AD and CH (Position 10).

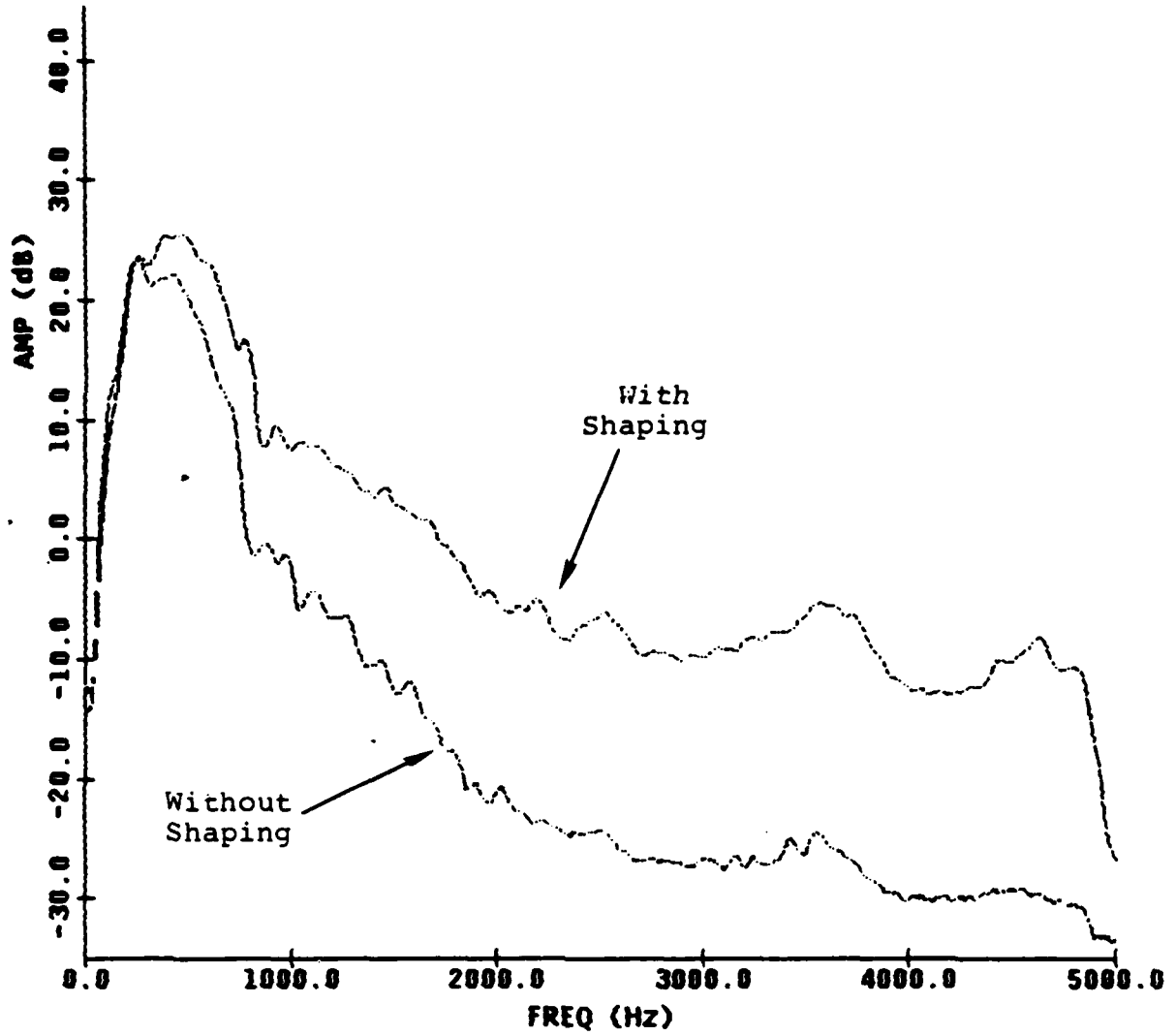


FIG. 8. Long-term speech spectra for speaker AD (Position 10).

Comparison of this plot with Fig. 19 of the multisensor project final report [1] shows that the "shaped" accelerometer signal spectrum bears a much greater resemblance to a pressure microphone signal spectrum than does the "unshaped" spectrum. Thus, one would expect the "shaped" accelerometer signal to sound more "natural" than the "unshaped" signal.

To evaluate the effect of the above spectral shaping method, we performed informal speech quality tests on the accelerometer signal and on the two-sensor signal (we used the first-order gradient microphone M12 and speaker KK's data), with and without spectral shaping and in 100 dB, 107 dB, and 114 dB levels of cockpit noise. At all three noise levels, the quality of the accelerometer signal alone was improved significantly by shaping; the shaping function found for speaker AD was used. The two-sensor signal with shaping showed a slight but audible improvement in quality over the unshaped case.

4. A TWO-MICROPHONE CONFIGURATION

As part of our long-term spectral analysis work, we compared the long-term speech spectrum with the long-term cockpit noise spectrum for each of our three gradient microphones. This study showed that Vought provides a higher signal-to-noise ratio (SNR) at low frequencies than M12 and EV 985, and that both M12 and EV 985 provide a higher SNR at high frequencies than Vought, with the high-frequency SNR being significantly higher for M12 than for EV 985. Fig. 9(a) shows the long-term speech and noise spectra for Vought located in front of the lips (position 7, see Fig. 1), and Fig. 9(b) shows the corresponding spectra for M12 located 45 degrees to one side of the mouth (position 3). (The speech spectra shown are peak speech spectra since we had added 12 dB to the RMS spectra, and the noise spectra correspond to the responses of the microphones to cockpit noise at about 102 dB SPL.) Figs. 9(a) and 9(b) show clearly the superiority (in terms of the SNR performance) of Vought at low frequencies and the superiority of M12 at high frequencies. The reason for the superior high-frequency response of M12 is that we located the microphone array substantially closer to the mouth than we could locate the bulkier Vought (see Table 3).

More important, the above result provided the basis for developing a two-microphone configuration in which a lowpass filtered Vought microphone signal is combined with a highpass filtered and amplitude adjusted M12 signal, with the same cutoff frequencies for the two filters. We performed the amplitude adjustment to make the high-frequency energy of the M12 signal equal to that of the Vought signal, which ensured the proper energy balance in the spectrum. We determined the value of the cutoff frequency, using an exhaustive procedure, as that which maximized the overall SNR (or alternatively the overall articulation index) of the two-microphone configuration. Fig. 10(b) shows the long-term speech and noise spectra for the two-microphone signal, with the cutoff frequency indicated by the dashed vertical line. For comparison, the spectra for Vought are shown in Fig. 10(a). The overall SNR of the two-microphone configuration was increased in this example by about 2.5 dB as compared to the Vought microphone alone. The articulation index score (which is a measure of speech intelligibility) was also increased significantly.

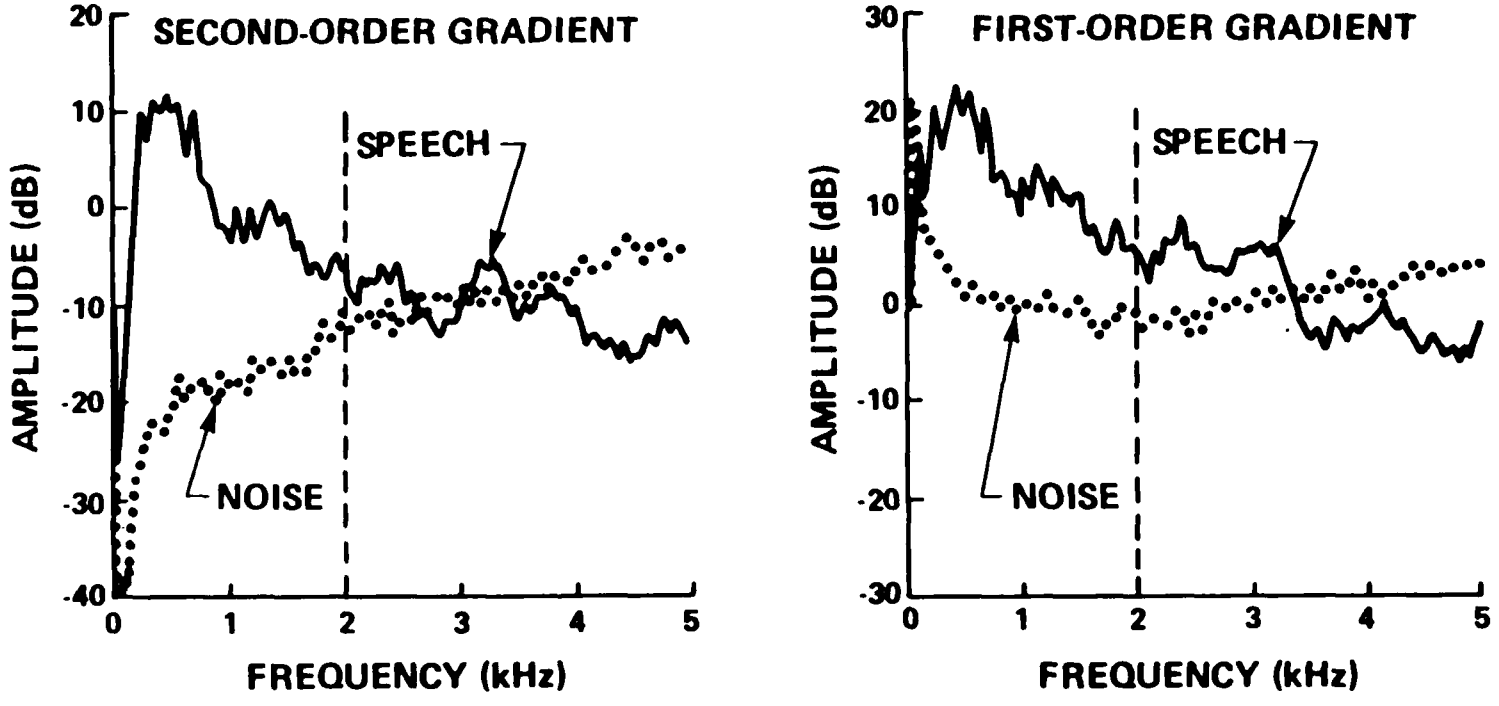


FIG. 9. Comparison of speech and noise spectra for (a) Vought second-order gradient microphone and (b) first-order gradient signal M12.

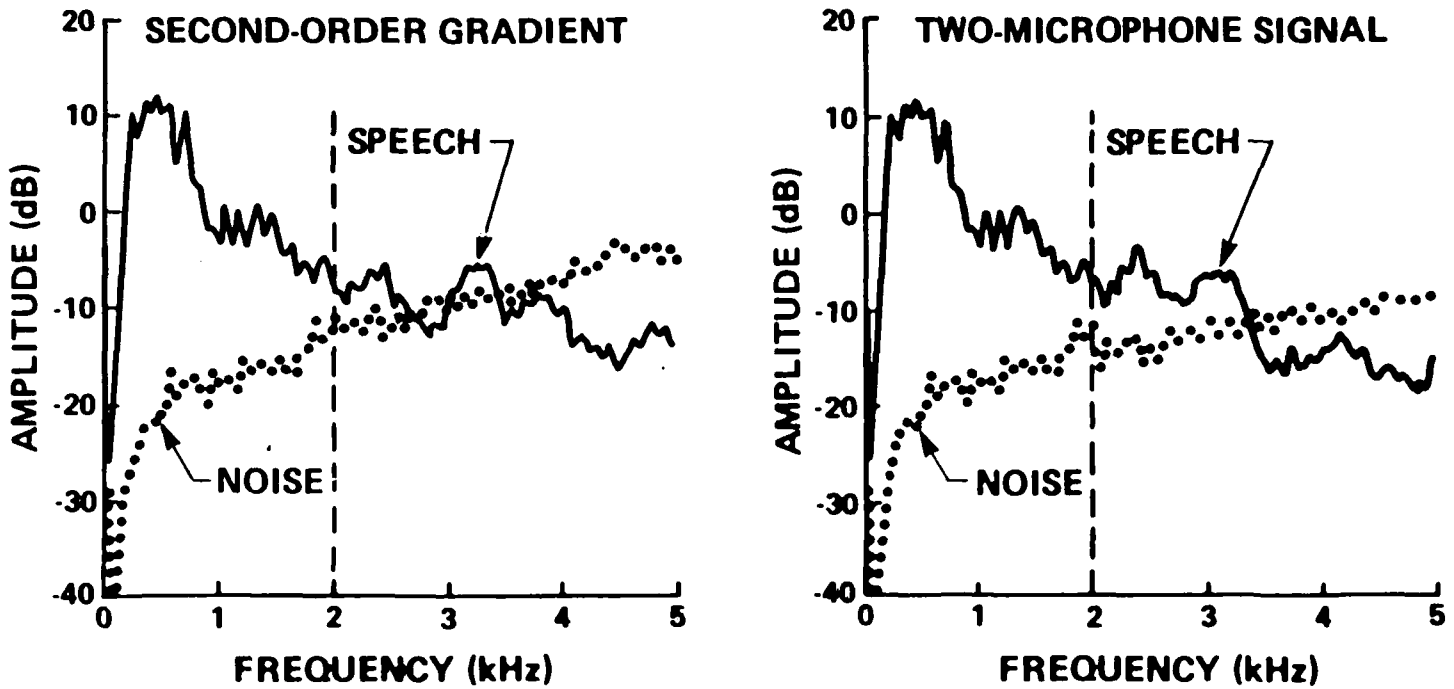


FIG. 10. Comparison of speech and noise spectra for (a) Vought second-order gradient microphone and (b) two-microphone system.

Fig. 11 shows the SNR of the two-microphone signal as a function of the cutoff frequency, with the optimal cutoff frequency indicated by the dashed vertical line. Notice from the figure that the value of the SNR at the cutoff frequency of 0 Hz is the SNR for M12 and the value at 5 kHz is the SNR for Vought. In computing the SNR, we assumed raised voice; in other words, we added 6 dB to the RMS spectra of speech recorded in the quiet background. The SNR plot shown in the figure exhibits a rather broad maximum, indicating a negligible sensitivity of the SNR to changes in the cutoff frequency from its optimal value; this property explains why we obtained, in informal listening tests, good speech quality and intelligibility, using the same cutoff frequency for different speakers (see Section 6.2). To improve the SNR of the two-microphone signal further, we incorporated the amplitude adjustment step mentioned in the preceding paragraph within the process of optimizing the cutoff frequency. With this modification, the SNR of the two-microphone signal was increased to 3.0 dB, as compared to -3.1 dB for M12 and -0.7 dB for Vought; all SNR's were computed for raised voice and 102 dB noise.

The two-microphone system just described is a desirable alternative to the earlier two-sensor system since the accelerometer used in the latter can pick up unwanted vibrations caused, for example, by the movement of the talker.

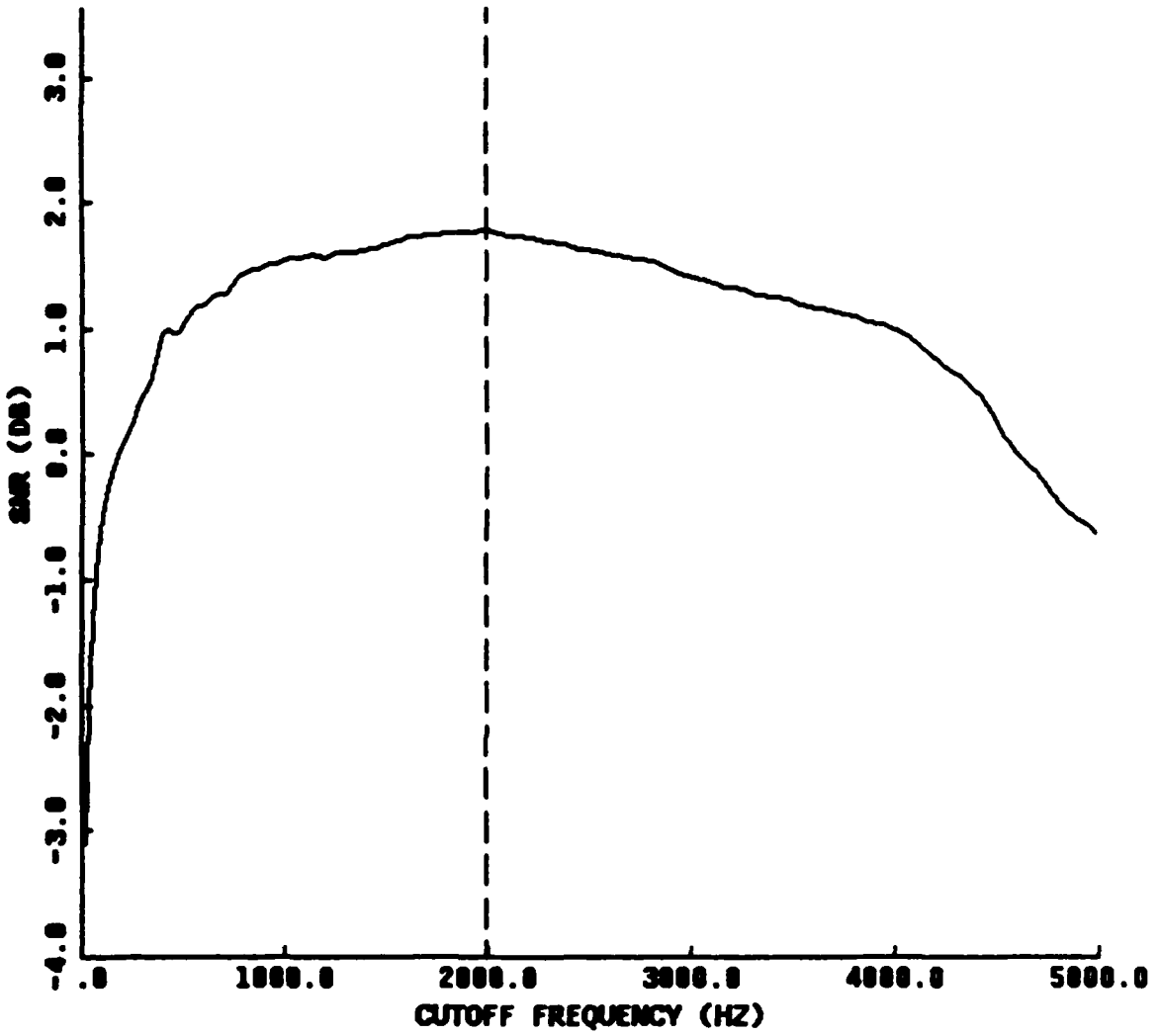


FIG. 11. Signal-to-noise ratio of the two-microphone signal plotted as a function of the cutoff frequency.

5. ADDITIONAL MULTISENSOR CONFIGURATIONS

Results from short-term spectral analysis of the sound-field data that we measured in the previous project indicated that the spectral amplitudes of nasal murmur (in relation to an adjacent vowel) were substantially weaker for all three gradient microphones we used than for a pressure microphone [1]. Therefore, we investigated the use of a nasal microphone as part of a multisensor configuration. Using the EV 985 for this purpose, we found that the best way of locating it is to place it directly under one nostril and not use a puffscreen, which allows the microphone to be very close to the sound source. Initial experiments in simulated noise yielded two results: 1) It is necessary to highpass filter the nasal microphone signal at about 500 Hz prior to combining with a lip microphone signal to reduce the level of the perceived noise; and 2) the (M12, EV 985) combination, with EV 985 being the nasal microphone, produced a slight improvement in the transduction of nasal sounds over M12 alone.

The results presented above suggested the following configurations involving three or four sensors: 1) (Vought, M12, and EV 985); 2) (accelerometer, Vought, M12, and EV 985); and 3) (accelerometer, Vought, and M12). In cases (1) and (2), EV 985 was included as a nasal microphone. In cases (2) and (3), the Vought signal had to be bandpass filtered before combining with the other sensor signals so that the accelerometer contributed information primarily to the low band; the Vought, to the middle band; and M12, to the high band.

6. FORMAL SUBJECTIVE TESTING IN NOISE

To compare the performance of the various multisensor configurations in broad-band noise, we simulated F-15 fighter aircraft cockpit noise at 95 dB and 115 dB SPL, and conducted a two-speaker Diagnostic Rhyme Test (DRT) [5] for speech intelligibility evaluation and a 10-point rating test for speech quality evaluation [1]. The cockpit noise was simulated in a reverberant room to ensure a stable, diffuse field, using the procedure described in [1].

6.1 Sensor Positions and Configurations

For each of the sensors, we determined the best position based on the results from articulation index analysis, short-term spectral analysis, and puffnoise study in which we evaluated, by listening, the level of the puffnoise picked up by a microphone with and without a puffscreen. Positions directly in front of the lips (positions 1 and 7; see Fig. 1) often led to objectionable levels of puffnoise, even though they produced high articulation index scores. Brief tests indicated that the physical placement of the three microphones in front of a talker would not affect detrimentally, because of interference, any of the microphone signals in the quiet or in noise. The sensor positions we chose are as follows: one side of the throat just above the Adam's apple (position 10) for accelerometer; 45 degrees to one side of the mouth (position 3) for M12; 45 degrees to the other side of the mouth for Vought; directly under one nostril for EV 985, as we were using this exclusively as a nasal microphone. Because the level of puffnoise was objectionable for M12 and Vought without puffscreens, we decided to use puffscreens for these two microphones; however, for both microphones we made the puffscreens as thin as possible to allow the closest possible placement to the mouth. As we noted in Section 5, EV 985 should be used without a puffscreen to allow closer placement to the nose for improved transduction of nasals.

From the results presented in Sections 3-5, we chose to evaluate the following seven multisensor configurations, involving four different sensors:

C1) Accelerometer and Vought
C2) Accelerometer and M12
C3) Accelerometer with shaping and M12
C4) Vought and M12
C5) Accelerometer, Vought, and M12
C6) Vought, M12, and EV 985
C7) Accelerometer, Vought, M12, and EV 985.

For mounting the multiple microphones, we developed a headpiece from a face shield by attaching metal rods that are adjustable both vertically and horizontally.

6.2 Multichannel Recording of Test Data

To guarantee identical speaking conditions for the seven multisensor systems, we mounted all four sensors simultaneously, recorded the sensor signals on a multichannel tape recorder, and combined afterwards the individual sensor signals to obtain the multisensor test data. We used two speakers, one male (RS) and one female (CH). Three ambient conditions were considered: quiet or no noise, 95 dB (cockpit) noise, and 115 dB noise. As test speech materials, we used the following items (refer to Appendix A for a listing of items 2-4):

1. The standard 232-word DRT material, with a different word list used for each speaker-noise condition [5]. A Radio Shack TRS-80 Model 100 personal computer was used to prompt each speaker for a given DRT list. Six lists, one for each speaker-ambient noise combination, were scrolled up the display screen, one DRT pair at a time, at a rate of one word pair every 1.3 seconds.
2. The same six sentences included in the multisensor project's database, to be used for subjective speech quality evaluation.
3. The 20-word TI vocabulary, containing the ten digits and ten command words, to be used in tests with a commercial speech recognition unit. For training data,

each speaker repeated the list 10 times, with a four-second interval between words. The same list was then repeated 20 times with two-second word spacing for use as test data.

4. A list of 44 minimal pair words, to be used in the study of the parallel-input system. The list was repeated 20 times, with roughly 1.5 seconds between words. In order to reduce the amount of data taken in the 115 dB noise, this section of data was omitted for that noise condition; we could simulate the data for the 115 dB noise environment digitally, using the minimal pair data taken in the quiet and using the recording of the noise alone.

The minimal pair words we chose are listed in Table 4 under seven categories that reflect one way we might use them in our investigation of the parallel-input multisensor system. From Table 4, we note that the category vowels contains high and low vowels and front and back vowels. For the category place for stops, we have two sets of words with [b, d, g] in initial position, one set with [b, d, g] in final position, one set with [p, t, k] in initial position, and one set with [p, t, k] in final position. In the place for nasals category, we have two sets of words with [m, n] in initial position and one set with [m, n, ng] in final position. The category place for fricatives has one set of words with [s, f] in initial position, one set with [s, f] in final position, and one set with [s, sh] in initial position. For the voiced-voiceless category, we have words with [t, d] in initial position, [t, d] in final position, [p, b] in initial position, [k, g] in initial position, [s, z] in final position, and [f, v] in final position. For the nasal-nonnasal category we have [n, d, l] in initial position for two sets of words; also, we have words with each of the three sound pairs in the final position: [m, b], [n, d], and [ng, g]. Finally, the category sonorants contains one example for each of six sonorants (sounds with no bursts or noise) in initial position. To keep the total number of distinct words reasonably small (44 in our case), we chose the words in such a way that one word may appear under more than one category. For example, the word 'net' appears under the categories: place for nasals, nasal-nonnasal, and sonorants.

Vowels

heed, hid, head, had, hod, hud, hood

Place for Stops

bet, debt, get, bode, dode, goad,
sog, sod, sob, dote, dope, doak,
toad, code, pode

Place for Nasals

met, net, mode, node, sawn, psalm, song

Place for Fricatives

leaf, lease, sod, shod, sin, fin

Voiced-Voiceless

bet, bed, pet, bet, goad, code,
dode, toad, dose, doze, leaf, leave

Nasal-Nonnasal

met, bet, mode, bode, net, debt, let,
song, sog, sawn, sod, psalm, sob, node,
dode, load

Sonorants

wet, yet, let, ret, met, net.

Table 4. Minimal pair words grouped under seven categories.

For combining the sensor signals in each of the seven multisensor systems given in Section 6.1, the filter cutoff frequencies and the amplitude adjustments for individual sensor signals were "optimized" subjectively through informal listening tests involving three listeners; we considered the two cockpit noise conditions only. For each sensor signal, we used about the same filter cutoff frequency for both speakers and both noise conditions; the required amplitude adjustments, however, varied across speakers and noise conditions.

6.3 Screening Evaluation

We compared the seven multisensor systems through informal listening tests and limited DRT tests, scoring only half the DRT words and with only three listeners. From this evaluation, we made the following observations: 1) Nasal sounds were transduced well by the two-sensor configurations C1, C2, and C3, all involving the accelerometer; 2) the benefit provided by the nasal microphone EV 985 in configuration C7 over C5 was only marginal; 3) the effect of including the nasal microphone in C6 (as compared to C4) was mixed in that perception of nasal sounds was slightly improved and perception of sounds [g, b, v, d] was degraded; and 4) the improvement provided by C5 over C1 or C4 was only slight. Based on these results and our interest to keep the number of sensors in a configuration small, we decided to include only the four two-sensor configurations C1-C4 for formal DRT and quality tests. For comparison purposes, we included also the three individual sensor cases: shaped accelerometer, Vought, and M12.

6.4 Generation and Scoring of Test Tapes

We generated DRT test tapes for the above-mentioned seven sensor systems. The test conditions for each sensor system included the two speakers and the two cockpit noise conditions, 95 dB and 115 dB. The tapes were scored at RADC, Hanscom Air Force Base, MA, using a panel of 12 trained listeners. For speech quality tests, we dubbed the six-sentence sets

twice more to obtain a total of 84 sets (= 28 conditions x 3 repetitions); dubbing was done to obtain three quality judgments per set from each listener. The test tapes contained a randomized order of these 84 sets. We used six experienced listeners, who rated the overall quality of each six-sentence set on a 10-point scale, with 1 being the worst quality and 10 being the best quality.

6.5 Speech Intelligibility Test Results

6.5.1 Overall DRT Scores

Table 5 gives the single-speaker DRT scores, the standard errors of listener means (given within parentheses), and two-speaker average DRT scores, for the various sensor and noise combinations. For convenience, we use the abbreviation ACC to denote the accelerometer without spectral shaping and ACC* to denote the accelerometer with spectral shaping; also, we use the notation (A,B) to denote the two-sensor configuration consisting of the sensors A and B, with A providing primarily the low-frequency information and B, the high-frequency information. First, let us consider the two-speaker average DRT scores. Comparing the two gradient microphones M12 and Vought, we see from Table 5 that in 95 dB noise M12 produced a slightly higher DRT score; in 115 dB noise, however, Vought produced a noticeably higher DRT score. Comparing (ACC, M12) with (ACC*, M12), we see from the table that spectral shaping of the accelerometer lowered the DRT score in both noise levels. Each of the two-sensor configurations (ACC, M12) and (ACC, Vought) produced a substantial improvement in the DRT score over the respective gradient microphone in 115 dB noise, and produced essentially the same DRT score in 95 dB noise; this result is in agreement with the result reported in our multisensor project final report [1]. Notice that the two-sensor configurations involving the accelerometer produced huge improvements in the DRT score over the accelerometer alone, even in 115 dB noise. The two-microphone configuration (Vought, M12) produced good improvements in the DRT score over Vought and M12 in 115 dB noise, and produced essentially the same DRT score in 95 dB noise. Among the two-sensor

SENSOR	95 dB			115 dB		
	RS	CH	AVE.	RS	CH	AVE.
ACC*	81.3 (1.2)	82.8 (1.2)	82.0 (0.83)	74.9 (2.1)	83.6 (1.3)	79.2 (1.63)
M12	96.7 (0.8)	94.3 (0.9)	95.5 (0.64)	85.2 (1.5)	85.2 (2.1)	85.2 (1.24)
(VOUGHT)	96.6 (0.7)	91.8 (0.8)	94.2 (0.81)	87.6 (0.6)	88.7 (0.7)	88.1 (0.45)
(ACC, M12)	96.9 (0.5)	92.6 (1.1)	94.7 (0.81)	90.6 (0.7)	94.1 (0.6)	92.4 (0.63)
(ACC*, M12)	93.9 (0.4)	94.0 (0.5)	93.9 (0.30)	90.5 (1.2)	92.6 (0.5)	91.5 (0.70)
(ACC, VOUGHT)	96.1 (0.9)	93.6 (0.8)	94.9 (0.67)	90.1 (1.4)	92.6 (0.7)	91.3 (0.82)
(VOUGHT, M12)	96.7 (0.6)	92.8 (0.5)	94.8 (0.63)	91.2 (0.7)	90.2 (1.0)	90.7 (0.61)

Table 5. Overall DRT scores for the seven sensor systems, two speakers, and two noise conditions. Numbers given within parentheses are standard errors of listener means.

configurations, (ACC, M12) seems to have the best overall intelligibility, although the DRT score differences among them are not large.

To determine the statistical significance of the difference in the two-speaker DRT scores, we used the standard two-tailed t test as follows. We present the test for a general case so that it can be used for comparing DRT scores and for comparing speech quality scores given below in Section 6.6. Suppose M1 and M2 are mean test scores over N judgments for the two sensor systems being compared, and V1 and V2 are the corresponding unbiased estimates of variances. (Unbiased estimate means that we divide the sum of the squares of the deviations from the mean by N-1 instead of N.) The parameter t is then given by (M1-M2)/E, where E is the square root of (V1+V2)/N. The degrees of freedom df is 2N-2. The level P of statistical significance of the difference M1-M2 is determined by referring to a t distribution table with the computed values of t and df. We considered a difference to be statistically significant if $P < 0.05$.

For the two-speaker DRT scores, the parameter t is again (M1-M2)/E, where E is now the square root of the sum of the squares of the two standard errors SE1 and SE2; df is 30 since N=16 (2 speakers x 8 listeners). We investigated the statistical significance for all possible pairwise comparisons of the seven sensor systems given in Table 5. For the 95 dB case, comparisons of ACC* with each of the other six sensor systems are all extremely significant ($P < 0.000001$); the case (ACC*, M12) vs. M12 is significant at $P < 0.05$; all other cases are not significant. For the 115 dB case, only the comparisons between any two two-sensor systems are not significant. In particular, comparisons involving a two-sensor system and one of the constituent single sensors are all highly significant ($P < 0.002$ or better).

Next, we consider the individual-speaker DRT scores given in Table 5. Comparing the DRT scores for the two speakers, we see that for each of the two microphones M12 and Vought the DRT score was higher for RS than for CH in 95 dB noise; in 115 dB noise, however, the DRT score for CH was equal to or slightly higher than the DRT score for RS. For the accelerometer alone, the DRT score in 115 dB noise was substantially higher for CH than for RS; the difference between their scores in 95 dB noise was only slight. In 115 dB noise, we

used a higher relative gain for the accelerometer signal for CH than for RS in the two-sensor configurations. As a result, the accelerometer's contribution to the two-sensor systems in 115 dB noise was higher for CH than for RS. For example, while the DRT score for (ACC, M12) was 4.3 points lower for CH than for RS in 95 dB noise, in 115 dB noise it was, in fact, 3.5 points higher for CH than for RS. Finally, for the two-microphone configuration (Vought, M12), the difference in the DRT score between RS and CH decreased from 3.9 points in 95 dB noise to 1.0 point in 115 dB noise, with the score for RS being higher in both cases.

The DRT score for M12 in the quiet was 96.9 (SE=1.0) for RS and 98.3 (SE=0.7) for CH. From a comparison of these two scores with the scores for the speakers used in the standard DRT tests [5], we found that our two speakers fell somewhere in the upper end of the DRT speakers, with the latter being ordered in terms of their DRT scores.

6.5.2 Attribute DRT Scores

The DRT words allow the speech intelligibility to be evaluated, for diagnostic purposes, for each of the six attributes: voicing, nasality, sustention, sibilation, graveness, and compactness [5]. Each of the rhyming pairs associated with the voicing attribute contains one word with an initial voiced consonant (e.g., bean) and another with an initial unvoiced consonant (e.g., peen). Nasality involves discrimination between nasal and non-nasal sounds (e.g., meat vs. beat); sustention involves discrimination between stop sounds and non-stop sounds (e.g., bee vs. vee); sibilation involves discrimination between sounds with intense high-frequency energy and sounds with low high-frequency energy (e.g., zee vs. thee); graveness involves discrimination between labial and non-labial sounds (e.g., bid vs. did); compactness involves discrimination between compact sounds and non-compact sounds (e.g., key vs. tea). The DRT results at the attribute level may have implications concerning performance in isolated-word recognition.

Since the DRT scores showed a larger variation over the sensors in 115 dB noise than in 95 dB noise, we present below the attribute DRT scores only for the 115 dB case. Also, since the results were similar for the two speakers, we present the results for speaker CH only.

Table 6 gives the attribute DRT scores for ACC*, M12, (ACC, M12), and (ACC*, M12). For each attribute, the table gives the combined score and the scores for each of two cases; the two cases are frictional and non-frictional for voicing; grave (labial) and acute for nasality; and voiced and unvoiced for all other attributes. Comparing (ACC, M12) with M12, we find from Table 6 that the addition of the accelerometer to M12 improved voicing by 13.2 points, sustention by 19.5 points, and graveness by 14.1 points. Comparing (ACC, M12) with (ACC*, M12), we see that the major effect of accelerometer spectral shaping was a reduction of the graveness score by 10.2 points. It is interesting to point out that although the overall DRT scores were only slightly different for M12 and ACC* (85.2 vs. 83.6), the attribute DRT scores were quite different (a 14.7 point difference in voicing, a 13.2 point difference in sustention, a 17.2 point difference in sibilation, and a 14.8 point difference in graveness).

Table 7 gives the attribute DRT scores for ACC*, Vought, and (ACC, Vought). The addition of ACC to Vought in the two-sensor configuration primarily improved sustention (14.8 points) and compactness (3.9 points).

Table 8 gives the attribute DRT scores for Vought, M12, and (Vought, M12). The addition of Vought to M12 in the two-microphone configuration improved voicing by 13.2 points, sustention by 7 points, and graveness by 4.7 points.

6.6 Speech Quality Test Results

Table 9 gives the single-speaker mean speech quality ratings and the two-speaker average ratings for the various test conditions; for each test condition, we averaged the available 18 rating scores (6 subjects x 3 judgments). From the table, we see that the addition of the accelerometer in each of the two-sensor configurations (ACC, M12) and (ACC, Vought) improved the mean speech quality rating modestly in both noise conditions. Comparing (ACC, M12) with (ACC*, M12), we find that spectral shaping lowered the mean rating in both noise levels; the decrease in the rating was more for CH than for RS. The two-microphone configuration produced a slight speech quality improvement over the two constituent

ATTRIBUTE	SENSOR			
	ACC*	M12	(ACC, M12)	(ACC*, M12)
VOICING	96.9	85.2	98.4	97.7
Frictional	96.9	71.9	98.4	98.4
Non-Frictional	96.9	98.4	98.4	96.9
NASALITY	100.0	98.4	99.2	99.2
Grave	100.0	98.4	100.0	100.0
Acute	100.0	98.4	98.4	98.4
SUSTENTION	89.8	76.6	96.1	96.1
Voiced	95.3	60.9	95.3	95.3
Unvoiced	84.4	92.2	96.9	96.9
SIBILATION	76.6	93.8	96.9	97.7
Voiced	90.6	89.1	95.3	100.0
Unvoiced	62.5	98.4	98.4	95.3
GRAVENESS	63.3	78.1	92.2	82.0
Voiced	95.3	81.3	96.9	84.4
Unvoiced	31.3	75.0	87.5	79.7
COMPACTNESS	75.0	78.9	82.0	82.8
Voiced	89.1	81.3	87.5	93.8
Unvoiced	60.9	76.6	76.6	71.9
OVERALL DRT	83.6	85.2	94.1	92.6

Table 6. Attribute DRT scores for ACC*, M12, (ACC, M12), and (ACC*, M12) in 115 dB noise, for speaker CH.

ATTRIBUTE	SENSOR		
	ACC*	VOUGHT	(ACC, VOUGHT)
VOICING	96.9	96.1	97.7
Frictional	96.9	98.4	98.4
Non-Frictional	96.9	93.8	96.9
NASALITY	100.0	100.0	98.4
Grave	100.0	100.0	96.9
Acute	100.0	100.0	100.0
SUSTENTION	89.8	79.7	94.5
Voiced	95.3	78.1	95.3
Unvoiced	84.4	81.3	93.8
SIBILATION	76.6	96.9	99.2
Voiced	90.6	96.9	100.0
Unvoiced	62.5	96.9	98.4
GRAVENESS	63.3	80.5	82.8
Voiced	95.3	92.2	95.3
Unvoiced	31.3	68.8	70.3
COMPACTNESS	75.0	78.9	82.8
Voiced	89.1	87.5	87.5
Unvoiced	60.9	70.3	78.1
OVERALL DRT	83.6	88.7	92.6

Table 7. Attribute DRT scores for ACC*, Vought, and (ACC, Vought) in 115 dB noise, for speaker CH.

ATTRIBUTE	SENSOR		
	VOUGHT	M12	(VOUGHT, M12)
VOICING	96.1	85.2	98.4
Frictional	98.4	71.9	96.9
Non-Frictional	93.8	98.4	100.0
NASALITY	100.0	98.4	100.0
Grave	100.0	98.4	100.0
Acute	100.0	98.4	100.0
SUSTENTION	79.7	76.6	83.6
Voiced	78.1	60.9	76.6
Unvoiced	81.3	92.2	90.6
SIBILATION	96.9	93.8	94.5
Voiced	96.9	89.1	93.8
Unvoiced	96.9	98.4	95.3
GRAVENESS	80.5	78.1	82.8
Voiced	92.2	81.3	93.8
Unvoiced	68.8	75.0	71.9
COMPACTNESS	78.9	78.9	82.0
Voiced	87.5	81.3	92.2
Unvoiced	70.3	76.6	71.9
OVERALL	88.7	85.2	90.2

Table 8. Attribute DRT scores for Vought, M12, and (Vought, M12) in 115 dB noise, for speaker CH.

SENSOR	95 dB			115 dB		
	RS	CH	AVE.	RS	CH	AVE.
ACC*	3.0	5.4	4.2	3.7	4.2	4.0
M12	8.1	5.5	6.8	2.2	2.3	2.3
VOUGHT	7.6	5.8	6.7	3.0	2.5	2.8
(ACC, M12)	8.7	6.0	7.4	4.5	3.6	4.1
(ACC*, M12)	8.6	5.4	7.0	4.7	3.0	3.4
(ACC, VOUGHT)	8.7	6.3	7.5	3.9	3.5	3.7
(VOUGHT, M12)	8.5	5.7	7.2	3.4	2.8	3.1

Table 9. Mean speech quality ratings for the seven sensor systems, two speakers, and two noise conditions.

microphones in both noise levels. For all sensors except the accelerometer, the mean quality rating was higher for RS than for CH in both noise levels; the relationship was reversed for the accelerometer. In fact, in 115 dB noise the mean rating for CH was highest for the accelerometer. The quality rating scores given in Table 9 suggest that the listeners were primarily judging the level of the background noise in the test sentences.

For evaluating the statistical significance of the differences in the two-speaker speech quality scores, we used the standard two-tailed t test as described above in Section 6.5.1. For the present case, the number of judgments N is 36 (2 speakers x 6 listeners x 3 judgments per item); the degrees of freedom df is therefore 70. We investigated the statistical significance of comparisons of each two-sensor system with its constituent single sensors, requiring a significance level of $P < 0.05$. (We did not consider (ACC*, M12) in this study.) Referring to Table 9, we note that the observed speech quality improvements are significant only for two cases in 95 dB noise, (ACC, M12) vs. ACC* and (ACC, Vought) vs. ACC*, and for three cases in 115 dB noise, (ACC, M12) vs. M12, (Vought, M12) vs. M12, and (ACC, Vought) vs. Vought; of these five cases, the first three are extremely significant ($P < 0.000001$),, the fourth case is significant at $P < 0.001$, and the fifth case is significant at $P < 0.007$. Among the three two-sensor systems, only the case (ACC, M12) vs. (Vought, M12) in 115 dB is significant ($P < 0.02$).

7. RECOGNITION TESTS WITH THE VERBEX 4000

We tested and compared the performance of the various single-input two-sensor systems and the individual sensors in speaker-dependent, isolated-word speech recognition, using the commercial recognizer Verbex 4000 and three different vocabularies. Below, we describe the Verbex 4000 recognizer and present and discuss the test results for the three vocabularies.

7.1 Description of Verbex 4000 Speech Recognizer

The Verbex 4000 is a speaker-dependent connected-word recognition system that can be used for isolated-word recognition. Using the system requires two cartridges. The Master Cartridge specifies the vocabulary, the grammar (an isolated-word grammar in our case), and the training script that prompts the user in training mode. The User Cartridge is used to store the speaker-dependent templates for all words in the vocabulary, and it is required for the recognition phase. The performance evaluation of each multisensor configuration involves two phases: a training phase and a recognition phase.

During the training phase, the Verbex unit first "enrolls" each word in the vocabulary; a "seed" template for a vocabulary word is created after the unit accepts two tokens of the word played into the unit's microphone port. After enrollment is complete, the unit prompts for the vocabulary words in sequence. If the unit cannot successfully train a particular utterance, it will prompt the user to repeat the word. However, after the word is input again, the unit will go on to prompt for the next vocabulary word, even if the second training attempt was unsuccessful. Each training pass involves a single run-through of the entire vocabulary.

We note that training the recognizer directly in a high noise condition may often fail. In our work, we conducted experiments to determine a sequence of training passes for each noise condition that would prove as successful as possible during testing. Included in the training procedures we considered were the following: N passes in the noise only, M passes in the quiet followed by N passes in the noise, and a "staged" training sequence. (We considered different

values for M and N.) The "staged" sequence would start with training in the quiet and lead up to the desired noise condition via training in one or more intermediate noise levels.

The Verbex Voice Planner Software, which was supplied with the Model 4000 and was installed on our IBM-XT, gave us the capability to create new vocabularies and to back up training data from memory cartridges onto floppy disks. For backups, the Verbex unit is interfaced with one of the IBM-XT communications ports, and the contents of the cartridge of interest are read while the cartridge is plugged into the Verbex unit. The restoration of training data from floppy disk to cartridge is performed in a similar fashion.

7.2 Tests Using the 20-Word TI Vocabulary

7.2.1 Generation of Tapes for Recognition Tests

For testing with the 20-word TI vocabulary on the Verbex 4000, we generated tapes containing the training and test sequences for each system-speaker-ambient noise condition. For each condition, we recorded the following data onto tape from the multichannel master tapes (see Section 6.2): first, ten sets of the TI words recorded in the quiet, for the first stage of training with up to ten tokens per word; second, ten sets of the TI words recorded in the ambient noise condition being considered, for the second stage of training with up to ten tokens per word; and third, twenty sets of the TI words recorded in the desired ambient noise condition, to be used in testing. For the sake of consistency, we employed the same two-sensor system parameters, filter cutoff frequency and gain, for both the training data in the quiet and the training and test data in the noise. (We used the "optimal" parameters determined through listening tests; see Section 6.2.)

We also digitized two full sets of the TI words from each set of "quiet" data for use in the enrollment phase of training the device. Because the words on the tapes occurred sequentially rather than in groups of two, the data needed to be rearranged; this was most easily accomplished by digitizing the data, then playing out in any order desired.

7.2.2 Test Results

The results of our recognition tests using the TI vocabulary are given in Table 10 for speaker RS and in Table 11 for speaker CH. We note that, for most tests, a 5 kHz lowpass filter was applied to the signal before it was input to the recognizer. This filter was used to supplement the Model 4000's built-in anti-aliasing filter, which has a "soft" stopband roll-off rate.

Considering the test results for the 95 dB case, we note that for all sensor systems except ACC* we used 3 training passes in the quiet and 8 training passes in 95 dB noise. For ACC*, we did not train in the quiet and used 10 training passes in 95 dB. Recognition tests were performed always using the full 20 passes. For ACC* and two-sensor systems involving the accelerometer, we highpass filtered the input signal prior to applying it to the recognition unit, because the Verbex unit had trouble training successfully without such filtering. For ACC*, we used in addition a sharp lowpass filter with cutoff at 3.2 kHz and a gradual lowpass filter (24 dB per octave) with cutoff at 1.5 kHz; in the absence of the lowpass filters, the high-frequency noise in the ACC* signal (enhanced by the accelerometer spectral shaping) caused training problems. We note, however, that we did not attempt to find an "optimal" choice of filters to use. Referring to Table 10, we see that for speaker RS recognition accuracy in 95 dB was either 99.5% or 100% for both gradient microphones and all three two-sensor systems. For M12, we used the same training cartridge and tested twice, obtaining 100% accuracy the first time and 99.5% accuracy (2 substitution errors) the second time; the small difference in accuracy between the two tests is not significant. The accelerometer alone produced 97.75% accuracy for RS. Considering the 95 dB case for speaker CH, we see from Table 11 that Vought yielded 100% accuracy. For (Vought, M12) we performed two separate runs of training and testing, and obtained 98.25% and 99.25%. The two-sensor system (ACC, Vought) had trouble training the words SIX, REPEAT, and EIGHT the required 8 times; however, we trained these words with several training attempts in the single-word training mode. We obtained a recognition accuracy of 93.75%; of the total 6.25% error, 5% was due to REPEAT. For (ACC, M12), additional training attempts were required for the word REPEAT; two separate runs of training and testing yielded 96.25% and 94.5%. M12 had

SENSOR SYSTEM	FILTERS USED	95 dB		115 dB	
		TRAINING SEQUENCE	RECOGNITION ACCURACY	TRAINING SEQUENCE	RECOGNITION ACCURACY
ACC*	400 Hz HPF 1.5 kHz LPF 3.2 kHz LPF	(0, 10)	97.75%	(0, 0, 10) (0, 10, 10) (0, 10, 0)	94% 98.75% 93.25%
M12	None	(3, 8)	100%, 99.5% (2 Tests)		
VOUGHT	5 kHz LPF	(3, 8)	99.5%	(3, 4, 5) Training Problems	17.5%
(ACC, M12)	200 Hz HPF 5 kHz LPF	(3, 8)	99.5%	(3, 5, ?)	Won't train in 115 dB
(ACC, VOUGHT)	200 Hz HPF 5 kHz LPF	(3, 8)	99.5%		
(VOUGHT, M12)	None	(3, 8)	100%	(3, 4, 5)	53.25%, 55.75%, 58.25% (3 Tests)

Table 10. Verbex 4000 test results for speaker RS, for the TI vocabulary. Training sequence is given in terms of number of passes used in quiet, 95 dB, and 115 dB in that order. In two cases, M12 in 95 dB and (Vought, M12) in 115 dB, we performed multiple tests using the same User Cartridge.

SENSOR SYSTEM	FILTERS USED	95 dB		115 dB	
		TRAINING SEQUENCE	RECOGNITION ACCURACY	TRAINING SEQUENCE	RECOGNITION ACCURACY
ACC*	400 Hz HPF 1.5 kHz LPF 3.2 kHz LPF	(0, 10)	96.75%	(0, 0, 10) (0, 10, 10) (0, 10, 0)	97.75% 97.5% 62%
M12	5 kHz LPF	(3, 8)	Training Problems		
VOUGHT	5 kHz LPF	(3, 8)	100%		
(ACC, M12)	200 Hz HPF 5 kHz LPF	(3, 8) (3, 8) (Two Runs)	96.25% 94.5%	(3, 8, ?)	Won't train in 115 dB
(ACC, VOUGHT)	200 Hz HPF 5 kHz LPF	(3, 8)	93.75%		
(VOUGHT, M12)	5 kHz LPF	(3, 8) (3, 8) (Two Runs)	98.25% 99.25%	(3, 8, ?)	Won't train in 115 dB

Table 11. Verbex 4000 test results for speaker CH, for the TI vocabulary. Training sequence is given in terms of number of passes used in quiet, 95 dB, and 115 dB in that order. In two cases, (ACC, M12) and (Vought, M12) in 95 dB, we performed two separate runs of training and testing.

difficulty training almost half of the vocabulary words; we therefore did not perform a recognition test for M12, for CH. The training problems with M12 and (ACC, M12) for CH may be attributed, at least in part, to the fact that some tape saturation occurred during data collection because of excessive recording gain. The accelerometer alone produced 96.75% for CH.

Next, let us discuss the results for the 115 dB case. For speaker RS, the two-microphone system did not train at all on the four words SIX, ENTER, REPEAT, and TWO; it trained successfully on all other words. Using the resulting training cartridge, we tested three separate times with the results of 53.25%, 55.75%, and 58.25%. The Vought microphone for speaker RS had many training problems and yielded only 17.5% accuracy. The other two-sensor systems we considered for either speaker did not train one-half or more of the vocabulary words in 115 dB. The accelerometer alone produced very good results in 115 dB. With 10 training passes directly in 115 dB (no training in the quiet or in 95 dB), ACC* yielded 94% for RS and 97.75% for CH. With 10 training passes each in 95 dB and in 115 dB, ACC* yielded 98.75% for RS and 97.5% for CH. The staged training therefore produced a sizeable performance improvement for RS and no significant change for CH. With 10 training passes in 95 dB only (none in 115 dB), ACC* produced 93.25% accuracy for RS and only 62% for CH. This experiment was an attempt to examine the effect on recognition performance caused by differences between training and test conditions. ACC* seems to have robust performance for RS but not for CH. The reason for this result may be that while CH increased her speaking level substantially in 115 dB relative to her level in 95 dB, RS increased his level to a lesser degree.

A word of caution about interpreting the recognition performance of ACC* reported above. While the accelerometer performance was impressive in 115 dB for the TI 20-word vocabulary, its performance for other vocabularies could be much lower, since the accelerometer is relatively insensitive to unvoiced sounds (see Sections 7.3 and 7.4).

In an attempt to improve the training and recognition performance of the Verbex unit in 115 dB noise, we investigated the effect of bandlimiting its input signal substantially below 5

kHz for each of the two sensor systems, (Vought, M12) and (ACC, M12); such bandlimiting renders the signal less noisy, as the noise amplitudes dominate the signal amplitudes at high frequencies. For (Vought, M12) in 115 dB noise with speaker RS, we performed several experiments to determine if lowpass-filtering the two-sensor signal would improve the recognition accuracy. Cutoff frequencies of roughly 2 kHz to 3.5 kHz were of interest; a 5 kHz cutoff frequency was used in our previous experiments in 115 dB noise. Because a filter with a cutoff frequency below about 3.2 kHz caused enrollment difficulties with some words, all enrollment was performed using a 5 kHz bandwidth. Training (3 passes in the quiet and 10 in 95 dB noise) was performed using a cutoff frequency in the prescribed range. We then attempted to train the system in 115 dB noise using this same cutoff frequency, and compared the resulting training performance with that found in the earlier (Vought, M12) experiments. We found that for the two cutoff frequencies we investigated, 2.2 kHz and 3.2 kHz, 9 words or more per training pass in 115 dB noise would not train. No recognition tests were performed because of the inferior training performance. We conclude from this investigation that bandlimiting (Vought, M12) in 115 dB noise in this fashion does not improve the Verbex unit's performance with this data.

We then attempted to repeat the original (Vought, M12) training procedure (3 passes in the quiet, 4 in 95 dB, and 5 in 115 dB, with a 5 kHz lowpass filter) and found that, in the first training pass for 115 dB, 11 words would not train. This performance was inconsistent with the performance observed previously, but was roughly equivalent to the performance of the more severely bandlimited signal discussed above. We suspect that 115 dB is too much noise for the Verbex unit to provide stable and consistent performance for all sensors but the accelerometer.

We also investigated the effect of more severely bandlimiting (through lowpass filtering) the (ACC, M12) mix on its training and test performance. As with the (Vought, M12) system, our filtering rationale was based on the idea that filtering out the signal's high-frequency noise should improve its recognition performance; however, because the low-frequency component of the (ACC, M12) signal is highly noise-resistant, we expected that such lowpass filtering would be more effective with this system than it was with (Vought, M12). As a first test, we

trained RS's (ACC, M12) using a 200 Hz highpass filter and a 3.2 kHz lowpass filter in the input line. Three training passes in the quiet and 5 in 95 dB were performed. In the first training pass in 115 dB noise, 10 words would not train; therefore, little performance benefit was realized with a lowpass cutoff frequency of 3.2 kHz. We then reasoned that, employing the same filters for (ACC, M12) that we used to test RS's ACC* in 115 dB noise (a highpass filter with a 400 Hz cutoff frequency cascaded with two lowpass filters with cutoff frequencies of 1.5 kHz and 3.2 kHz, respectively), similar results to those for ACC* should be achieved. It is important to note that the 1.5 kHz filter used was a "soft" filter with a 24 dB/octave rolloff rate; therefore, a significant component of the M12 signal could still be heard in the (ACC, M12) mix after filtering. Using these filters, we performed 5 training passes in 95 dB and 10 in 115 dB; because the signal was fairly quiet due to the accelerometer's large share in the mix, no training was needed in the quiet. Training in 115 dB went smoothly. Recognition accuracy for 20 passes of test data in 115 dB was 95.75%, which is within the range of test results found for RS's ACC* (94%, 98.75%, and 93.25%). Keeping in mind that for RS in 95 dB, (ACC, M12) with a lowpass cutoff frequency of 5 kHz yielded a recognition accuracy of 99.5%, we can conclude that varying the filter cutoff frequencies as the background noise level changes allows the (ACC, M12) system to be used successfully for speech recognition in noise levels as high as 115 dB.

7.2.3 Conclusions

Since the recorded M12 signal for speaker CH involved tape saturation due to excessive recording gain, we used primarily the recognition test results for speaker RS (Table 10) in making the conclusions given below. In 95 dB noise, the two gradient microphones and the two two-sensor systems provided similar recognition accuracy (about 99.5% for RS), which was modestly higher (by about 2% for RS) than the accuracy provided by the accelerometer. In 115 dB noise, only the accelerometer signal trained and tested successfully, yielding a recognition accuracy of about 98%. As we expected, the recognition accuracy for the accelerometer changed only slightly from 95 dB to 115 dB (accuracy increased by 1% in 115 dB!), which indicates its robustness in noise.

The recognition test results for the TI vocabulary suggest the use of the two sensors ACC and M12 in one of two ways, for achieving the best performance for the TI vocabulary. The first method involves a switching strategy based on the noise level: Use M12 for low noise and ACC for high noise. (We conjecture that the threshold noise level for switching is about 100 dB.) The second method uses the two-sensor system (ACC, M12) and filters the output to bandlimit the two-sensor signal as follows: 400 Hz HPF and 5 kHz LPF for low noise and 400 Hz HPF, 1.5 kHz LPF (gradual cutoff), and 3.2 kHz LPF (sharp cutoff) for high noise.

7.3 Tests Using a 25-Word Minimal Pairs Vocabulary

7.3.1 Selection of the Vocabulary

The recognition test results reported above do not provide any conclusive proof for the superiority of the two-sensor systems over the individual microphones; this is the case because 1) the 20-word TI vocabulary is too easy at 95 dB for all our sensor systems in the sense that each sensor system produced near 100% recognition accuracy for RS and 2) 115 dB is too much noise for all but the accelerometer to even train. We had two options available for resolving the problem: 1) test the TI vocabulary at an intermediate noise level such as 100 dB, and 2) use a more difficult vocabulary in 95 dB noise. We rejected the first option because the accelerometer would have continued to produce a high recognition accuracy (mid to upper 90's) for the TI vocabulary, thus leaving little or no room for improvement with a two-sensor system consisting of an accelerometer and a gradient microphone. For the second option, we used a subset of our minimal pairs database in 95 dB noise, which we expected to provide ample room for improvement with a two-sensor system. To gain some insights into recognition performance with minimal pair words and to guide our selection of a proper subset of minimal pair words, we examined in detail the attribute scores of the DRT tests in 95 dB noise, for speaker RS. We note that the rhyming word pairs used in the DRT are, in fact, minimal pairs because each pair of words differ only in the initial consonant.

Table 12 gives the DRT attribute scores for the individual sensors ACC*, M12, and

ATTRIBUTE	A*	M	V	(A, M)	(A, V)	(V, M)
VOICING	96.1	100.0	100.0	100.0	98.4	98.4
Frictional	95.3	100.0	100.0	100.0	96.9	96.9
Non-Frictional	96.9	100.0	100.0	100.0	100.0	100.0
NASALITY	86.7	95.3	94.5	93.8	93.0	97.7
Grave	78.1	92.2	89.1	87.5	85.9	95.3
Acute	95.1	98.4	100.0	100.0	100.0	100.0
SUSTENTION	81.3	96.1	93.8	94.5	96.1	96.1
Voiced	92.2	98.4	98.4	95.3	100.0	100.0
Unvoiced	70.3	93.8	89.1	93.8	92.2	92.2
SIBILATION	75.0	100.0	100.0	100.0	99.2	100.0
Voiced	82.8	100.0	100.0	100.0	100.0	100.0
Unvoiced	67.2	100.0	100.0	100.0	98.4	100.0
GRAVENESS	74.2	90.6	94.5	94.5	91.4	90.6
Voiced	98.4	100.0	100.0	98.4	98.4	100.0
Unvoiced	50.0	81.3	89.1	90.6	84.4	81.3
COMPACTNESS	74.2	98.4	96.9	98.4	98.4	97.7
Voiced	92.2	100.0	96.9	100.0	100.0	98.4
Unvoiced	56.3	96.9	96.9	96.9	96.9	96.9
OVERALL DRT	81.3	96.7	96.6	96.9	96.1	96.7

Table 12. DRT attribute scores for various sensors in 95 dB noise, for speaker RS. Symbols A, A*, M, and V are used to denote unshaped accelerometer, shaped accelerometer, M12, and Vought, respectively.

Vought and for the two-sensor systems (ACC, M12), (ACC, Vought), and (Vought, M12), for speaker RS in 95 dB noise. We see from the table that the attributes nasality-grave, sustention-unvoiced, and graveness-unvoiced seemed to be a problem (as indicated by lower scores) for all sensors. For the accelerometer, only voicing (both frictional and non-frictional) and nasality-acute were not a problem. Even though the accelerometer produced 97.75% accuracy for the 20-word TI vocabulary in 95 dB, it would give only 50% accuracy for a vocabulary of words characterized by the graveness-unvoiced attribute.

We note that the DRT test results we received from Hanscom AFB provided, for each attribute, the scores for the two cases: attribute present (i.e., the correct word has the attribute in question) and attribute absent (i.e., the correct word does not have the attribute). For items in Table 12 that have lower DRT attribute scores, we examined the scores for the attribute present and attribute absent cases. This data is given Table 13, where we placed asterisks next to scores that are lower than 92%, indicating that these cases provide room for improvement. From Table 13, we see that the accelerometer yielded only 31% correct recognition (by human listeners) when the spoken words with initial unvoiced consonants that were not compact were compared against rhyming words whose initial consonants were compact (e.g., tea vs. key, so vs. show, peg vs. keg, fit vs. hit, etc.). From the results given in Table 13 and from the list of DRT words, we found that the following sound pairs led to lower DRT attribute scores for all our sensors in general: (m,b), (sh,ch), (th,t), (f,p), (w,r), (b,d), (m,n), (p,t), and (f,th).

When we attempted to use all 44 words of the minimal pairs database with the Verbex 4000, the unit indicated that the grammar complexity for this vocabulary was 119%. The recommended complexity is below 80%. We then found out that the complexity was 63% for the first 20 words of the minimal pairs database and 85% for the first 30 words. Therefore, we decided to use a vocabulary of 25 words.

From the 44-word minimal pairs database, we chose a subset of 25 words by covering all the sound pairs (mentioned above) that yielded lower DRT attribute scores. The subset we chose is given in Table 14. The chosen subset consisted of all words in the categories, place for

SENSOR	NASALITY GRAVE	SUSTENTION		SIBILATION		GRAVENESS UNVOICED	COMPACTNESS	
		Voiced	Unvoiced	Voiced	Unvoiced		Voiced	Unvoiced
ACC*	100, 56*	84*,100	78*, 63*	97, 69*	72*, 63*	50*, 50*	100, 84*	31*, 81*
M12	97, 88*		88*, 100			81*, 81*		
VOUGHT	91*, 88*		78*, 100			84*, 94		
(A, M)	100, 75*		88*, 100			97, 84*		
(A, V)	100, 72*		84*, 100			94, 75*		
(V, M)	100, 91*		84*, 100			88*, 75*		

Table 13. Problem areas for various sensors, as indicated by DRT attribute scores (less than 92%). The two scores given in each cell correspond to the two cases, attribute present and attribute absent, respectively (see text). Asterisks are used to indicate scores that are less than 92%.

stops, place for nasals, and nasal-nonnasal and some words from the categories, place for fricatives, voiced-voiceless, and sonorants. The grammar complexity of the chosen vocabulary, as reported by the Verbex Voice Planner Software, was 74%.

Pode,	Toad,	Dope,	Dote,	Doak,
Code,	Shod,	Sod,	Met,	Bet,
Mode,	Bode,	Psalm,	Sob	Debt,
Get,	Dode,	Goad,	Sog,	Net,
Node,	Sawn,	Song,	Let,	Load.

Table 14. A vocabulary of 25 minimal pair words.

7.3.2 Tests for Single Sensors

Table 15 lists the tests for single sensors that we performed for speaker RS using this vocabulary in 95 dB ambient noise. Because we had collected only 20 passes of minimal pairs data, we used 5 tokens per word for training and 15 tokens per word for testing. It should be noted that ACC here was unshaped. However, we believe that with the severe bandlimiting applied to the signal here, shaping the signal would have made no significant difference. From Table 15, we see that M12 and Vought yielded a substantially higher accuracy than did the accelerometer. In contrast, the differences among the recognition accuracies of the three single sensors were relatively small for the 20-word TI vocabulary in 95 dB noise. That the accuracies of the three sensors were lower for the minimal pairs vocabulary than for the TI vocabulary is a confirmation of the higher complexity of the minimal pairs vocabulary.

Tables 16 and 17 list the phoneme confusions that occurred in each test. The numbers in the TOTAL column of each table indicate the maximum number of confusions possible among words differing only in the specified phoneme pair. For example, the words BODE and PODE could yield a total of 30 errors; this would occur if all 15 tokens of BODE were heard as PODE, and vice-versa. Table 16 shows that ACC's greatest weaknesses for the test words are

SYSTEM	FILTERS	RECOGNITION ACCURACY
SINGLE SENSORS:		
ACC	400 Hz, 1.5 KHz LPF, 3.2 KHz LPF	61.9%
M12	5 KHz LPF	78.9%
VOUGHT	5 KHz LPF	82.1%
TWO-SENSOR SYSTEMS:		
(VOUGHT, M12)	5 KHz LPF	80.8%
(ACC, M12)	200 Hz HPF, 5 KHZ LPF	81.3%

Table 15. Recognition accuracies obtained using the Verbex 4000 on our 25-word minimal pairs vocabulary, for single sensors and two-sensor systems, for speaker RS in 95 dB noise.

CONFUSION CLASSES	TOTAL POSSIBLE	A	M	V	(A, M)	(V, M)
VOICING						
Non-Frictional						
[B], [P]	30	2	5	2	2	5
[D], [T]	90	4	0	2	1	0
[G], [K]	30	0	0	1	2	0
NASALITY						
Grave						
[M], [B]	90	6	7	3	5	4
Acute						
[N], [D]	60	1	11	6	2	16
GRAVENESS						
Voiced						
[B], [D]	60	7	1	3	7	5
[M], [N]	90	10	4	6	2	0
Unvoiced						
[P], [T]	60	10	2	0	4	2
COMPACTNESS						
Voiced						
[G], [D]	60	24	2	5	9	11
[G], [B]	60	1	9	8	11	10
Unvoiced						
[K], [T]	60	0	0	0	2	0
[K], [P]	60	12	23	10	10	4
[SH], [S]	30	11	0	0	0	0

Table 16. Phoneme confusions for Verbex tests of 25 minimal pair words, grouped by DRT category, for speaker RS in 95 dB noise. The numbers tabulated indicate the number of confusions. The abbreviations, A, M, and V denote, respectively, accelerometer, M12, and Vought.

CONFUSION CLASSES	TOTAL POSSIBLE	A	M	V	(A, M)	(V, M)
[B], [K]	30	0	0	0	1	0
[B], [L]	30	5	1	3	3	2
[B], [N]	60	0	0	1	0	0
[D], [K]	60	1	0	0	1	0
[D], [L]	30	1	0	0	0	0
[D], [M]	60	0	0	1	1	3
[D], [P]	60	0	0	2	0	2
[G], [K]	30	0	0	0	2	0
[G], [L]	30	2	0	1	0	0
[G], [M]	60	0	1	1	0	0
[G], [N]	60	0	2	2	0	3
[G], [P]	30	2	1	0	1	0
[G], [T]	30	0	0	0	1	0
[K], [L]	30	0	1	1	0	0
[K], [M]	30	1	0	0	0	1
[L], [M]	30	1	1	5	2	0
[L], [N]	30	2	0	0	0	0
[L], [P]	30	0	2	0	1	1
[M], [P]	30	3	2	0	0	0
[N], [NX]	30	0	3	3	1	3
[N], [T]	30	1	0	0	0	0
Total Number of Non-Minimal Word Confusions		36	1	1	1	0

Table 17. For Verbex tests of 25 minimal pair words, phoneme and word confusions that do not fall into DRT categories, for speaker RS in 95 dB noise. The numbers tabulated indicate the number of confusions. The abbreviations A, M, and V denote, respectively, accelerometer, M12, and Vought.

in the DRT categories of compactness and graveness, which were also problem areas for ACC* in the DRT tests for speaker RS in 95 dB (see Table 13). However, the phoneme confusions for the microphones were less consistent with the DRT test results than the confusions for ACC were. For example, both microphones showed problems in the compactness category for the Verbex tests, unlike for the DRT tests. Also, the problems encountered in the graveness unvoiced category for the DRT tests are not reflected in the Verbex test results for either microphone.

Several other items in Tables 16 and 17 are worth noting. First of all, the number of non-minimal word confusions, in which words differing by more than one phoneme were confused, was far greater for ACC than for any of the other systems tested. Most of these errors occurred where words were incorrectly recognized as ending in [ow] [d]. Secondly, although the tables do not indicate in which direction the errors occurred (i.e., [n] recognized as [d], or vice-versa), in a few cases the errors occurred much more frequently in one direction than in the other. For example, [g] was recognized as [d] in 21 of the 24 [g],[d] confusions that occurred for ACC. Also for ACC, [k] was recognized as [p] for all 12 of the [k],[p] confusions, and [s] was recognized as [sh] for 10 of the 11 [s],[sh] confusions. For Vought, [p] was recognized as [k] in 8 of the 10 [k],[p] confusions that occurred.

The test results for single sensors showed clearly that there was ample room for improvement because even the highest recognition accuracy was only 82.1%. Before we performed recognition tests for two-sensor systems, we examined the extent of improvement achievable under an ideal situation, as described below. We assumed the ideal situation in which the two-sensor system (S1, S2) would retain the good properties of both sensors S1 and S2, by yielding a substitution error only if both sensors yielded that error; in other words, the two-sensor system would not produce a substitution error if either sensor did not produce that error. From the confusion matrices of the single sensors, we computed the number of recognition errors for each two-sensor system, assuming the foregoing ideal condition. The resulting recognition accuracies were 90.4% for (Vought, M12), 93.3% for (ACC, M12), and 95.7% for (ACC, Vought), which represented substantial improvements over the performance of M12 and Vought. Encouraged by these potentially large performance improvements, we decided to test the two-sensor systems.

7.3.3 Tests for Two-Sensor Systems

The (Vought, M12) and (ACC, M12) signals for the 25 minimal pair words were generated digitally. For (Vought, M12), the Vought signal was lowpass-filtered at 1800 Hz and the M12 signal was highpass-filtered at the same frequency. The two resulting signals were then combined in the same proportion used for the analog mixing of the TI data. Similarly, for (ACC, M12), the M12 signal was highpass-filtered at 1500 Hz before the two signals were combined in the correct proportion. To reduce boominess, two 200-Hz analog highpass filters were applied to the (ACC, M12) mix during testing. All digital filters used were 39th order FIR filters.

Table 15 lists the overall results of the two tests. The phoneme confusions that occurred during the tests are listed in Tables 16 and 17. These results make it clear that the ideal situation that we suggested earlier did not apply here. For example, although the [k],[p] distinction was more successful for both mixes than it was for the individual sensor signals in each mix, the [g],[b] distinction for (ACC, M12) and the [g],[d] and [n],[d] distinctions for (Vought, M12) were less successful for the mixes than for the single sensor signals. Therefore, phoneme distinctions for the two-sensor systems were not always as good or better than the same distinctions for the single sensors. It is also interesting to note that for (Vought, M12), [g] was heard as [d] for 8 of the 11 [g],[d] confusions, and [n] was heard as [d] for all 16 of the [n],[d] confusions. Because the overall recognition accuracies for these tests showed no improvement over the results for the single sensor tests, we decided not to test (ACC, Vought) for the minimal pair words.

7.3.4 Additional Tests

We performed three sets of additional tests, using the Verbex unit on the 25-word minimal pairs vocabulary: 1) tests in the quiet; 2) training in the quiet and testing in 95 dB; and 3) tests in 95 dB with increased training. The results of these tests are presented below.

First, we trained and tested the Verbex unit in the quiet for M12 and Vought, to obtain

baseline data (for comparison with the test results in 95 dB, reported in Section 7.3.2) as well as to determine (via such comparisons) if 95 dB noise is a problem for the two microphones. Because we had a total of only 20 tokens per word, we trained the recognizer using five tokens, of which the final two were the same as the ones used for enrollment, and performed tests on the remaining 15 tokens, which produced the same test data size of 375 (=15 tokens x 25 words) as in the 95 dB tests described previously. The resulting recognition accuracy for the quiet condition was 65.1% for M12 and 68.5% for Vought. From Table 15, these figures were, respectively, 78.9% and 82.1%, for the 95 dB case. Therefore, the recognition accuracy decreased substantially in the quiet relative to the 95 dB case, a result we did not expect. When we discussed this issue with Dr. K. Ganesan of GTE Laboratories (who was formerly with Verbex), he suggested that we reduce the input level of the Verbex unit to below the recommended value of 300 mv (peak-to-peak). From his experience with the Verbex 3000, he observed that higher input levels could deteriorate the unit's performance. In our tests in 95 dB noise and in the quiet, the input level varied over a range of 150-600 mv. Following Dr. Ganesan's recommendation, we included an additional attenuator and maintained an input level below 300 mv. We repeated the test for M12 in the quiet, using five unique training tokens per word (i.e., we did not re-use the two enrollment tokens) and using the remaining 13 tokens per word for testing; this test yielded a recognition accuracy of 72.9%, which was significantly larger than the earlier result obtained with a larger input level. For comparison, we also repeated the test for M12 in 95 dB noise using the lower input level, with a training sequence of 3 tokens in the quiet followed by 5 tokens in 95 dB; this test yielded 76.3%, which was slightly lower than the earlier result of 78.9% given in Table 15. Although use of the lower input level narrowed the gap between the recognition accuracies in the quiet and in 95 dB noise, the accuracy was still higher in 95 dB noise. We offer the following three possible reasons for this unexpected result:

1. Speakers tend to talk more clearly and perhaps with less variability in (95 dB) noise than in the quiet.
2. Certain distance measures (e.g., log energy) are more sensitive to small variations in background interference in the quiet than in 95 dB noise.

3. Presence of some noise helps mask small speaking differences among the tokens of the same word.

In any case, since performance of the Verbex recognizer was not worse in 95 dB noise than in the quiet we concluded that 95 dB noise is not a problem for the microphones M12 and Vought.

Second, we trained the Verbex unit in the quiet (5 tokens per word) and tested it in 95 dB noise (15 tokens per word), for M12 and (ACC, M12); our motivation was to see if the two-sensor system offered an advantage in such a "cross-condition test". We used the same filters as given in Table 15. The recognition accuracy was 47.2% for M12 and 54.9% for (ACC, M12); in contrast, the figures for the case involving both training and testing in 95 dB noise were, respectively, 78.9% and 81.3%. These results show that the performance deterioration (going from training in 95 dB to training in the quiet) was less for (ACC, M12) than for M12. We note, however, that the cross-condition performance of both M12 and (ACC, M12) was unsatisfactory.

The third test we performed was based on the following observation. Because the microphones M12 and Vought transduce speech quite well at 95 dB (i.e., yield good SNR's), the training sequence with 3 tokens per word in the quiet and 5 in 95 dB, which we used for the 95 dB test results given in Table 15, might be roughly equivalent to training with 8 tokens per word, all in 95 dB. This effectively increased training might have caused the recognition accuracy in 95 dB to be better than that in the quiet condition (since the latter case involved only 5 training tokens). To pursue this issue further, we increased the number of training tokens per word to 3 in the quiet plus 10 in 95 dB noise and tested on the remaining 10 tokens, for Vought. The recognition accuracy was found to be 80.8%, which is not significantly different from the result for Vought given in Table 15.

The results presented above indicate that the low recognition accuracy of Vought and M12 in 95 dB noise was neither due to background noise nor due to lack of sufficient training; it was, we hypothesized, due to the duration of the initial consonant of each minimal pair word being considerably shorter than the duration of the following vowel and consonant. See below for more discussion of this point.

7.3.5 Conclusions

Several conclusions may be drawn from the above-reported recognition test results. First, using the accelerometer alone is, in general, not a good idea, since it yielded only 61.9% recognition accuracy for the 25-word minimal pairs vocabulary. (In contrast, the accelerometer yielded about 97% accuracy for the 20-word TI vocabulary.) As mentioned earlier, the low accuracy provided by the accelerometer is in general agreement with the DRT attribute scores presented earlier; we note that many of the minimal pairs in the 25-word vocabulary involved attributes for which the DRT scores were low for the accelerometer.

Second, the recognition accuracies in 95 dB noise for the two microphones and the two two-sensor systems ranged around 80% and were not significantly different from each other. This result is, in one sense, in agreement with the DRT test results since the DRT scores of the foregoing four sensor systems were also not significantly different from each other in 95 dB noise (see Table 5). There is, however, one important difference: While the human recognition accuracy, as given by the DRT tests, was about 95%, the machine recognition accuracy, as given by the tests on the Verbex unit, was only about 80%. The reason for the lower machine recognition accuracy, we hypothesized, was that the duration of the initial consonant in the minimal pair words was quite short relative to the duration of the following vowel and consonant. The initial consonants in each minimal pair (e.g., node and dode) were different, with the rest supposed to be identical in both words. Because the overall distance measure between a template and a test token is computed by averaging the frame-by-frame distances, it can easily happen that phonetic differences (as computed by the recognition device) in the short-duration initial consonant are masked by even small speaking differences in the long-duration vowel and consonant part; this masking effect, caused by the extreme durational differences between the two parts of the words, makes the machine recognition task difficult, thus yielding a lower recognition accuracy.

Third, since the recognition accuracy of the Verbex unit was not higher in the quiet (it was, in fact, lower) than in 95 dB noise, it is apparent that, as in the case of the 20-word TI vocabulary, acoustic background noise at 95 dB was not a problem for Vought and M12.

Fourth, the two-sensor system (ACC, M12) produced a significant improvement over M12 (54.9% vs. 47.2%) for the case where we trained the Verbex unit in the quiet and tested in 95 dB. Unfortunately, the improved performance was still unsatisfactory.

Finally, for effective demonstration of the advantage of a two-sensor system over its constituent gradient microphone for speech recognition, we inferred from the results reported above that we needed a vocabulary that 1) was more complex than the 20-word TI vocabulary (so that the accelerometer alone would not yield a high recognition accuracy), 2) kept durational problems as in the minimal pairs vocabulary to a minimum, and 3) had a noise level higher than 95 dB but below 115 dB (e.g., 105 dB).

7.4 Tests Using a 13-Word Minimal Pairs Vocabulary

7.4.1 Selection of the Vocabulary

For this vocabulary, we chose a 13-word subset of the minimal-pair words, listed in Table 18. To reduce the severity of the durational problem, all of the words differ from each other, at the least, by either the vowel (as in "heed" and "hud") or by both initial and final consonants (as in "sog" and "shod").

Pet,	Sog,	Heed,	Leave,	Hid,	Hood,	Head,
Node,	Dote,	Fin,	Had,	Hud,	Shod.	

Table 18. A vocabulary of 13 minimal pair words.

7.4.2 Generation of 105 dB Data

We digitally generated the 105 dB waveform files for this vocabulary as follows. For each utterance, we performed sample-by-sample addition of the digitized microphone (M12 or Vought) signal in 95 dB noise with an amplified version of the digitized noise-only responses of

the microphone in 95 dB noise. We set the noise amplification factor to be 9.54 dB, which yielded a noise level of 105 dB in the resulting waveform. We broke up the noise response of the microphone into 20 segments of about.1.5 seconds each, and used a different noise segment with each of the 20 tokens of a given word. The foregoing digital simulation assumes that the talker's speaking level remains the same in 95 dB noise and in 105 dB noise. (If, in fact, the talker raised his speaking level by, say, 3 dB in going from 95 dB noise to 105 dB noise, then the waveform we generated, as described above, would correspond to a background noise level of 108 dB.) As a quick check of the noise level in the new waveform files, for each microphone, we compared a plot of R0 (energy) for a 95 dB file with a plot of R0 for the corresponding 105 dB file. The energy difference between the two cases during non-speech regions, which is a measure of the difference in the two noise levels, averaged between 10 and 11 dB for Vought and between 9 and 10 dB for M12.

7.4.3 Test Results

The recognition accuracies for the tests of the 13-word vocabulary for speaker RS are shown in Table 19. The tests in 95 dB were performed for reference purposes. We tested the accelerometer (ACC) first, to ensure that its recognition accuracy allowed sufficient room for improvement. Because ACC is essentially insensitive to acoustic noise, we assumed that the ACC signal was the same for both 95 and 105 dB noise levels.

Some inconsistencies in these test results are apparent. The two recognition accuracies shown for (ACC, M12) in 105 dB differ by 34.3%. The procedures followed for these two tests differed only in the noise response waveform file used in each case; this file was played at the beginning of each training and test session so that the Verbex unit could estimate the characteristics of the background noise. During data collection we recorded this noise response for each sensor before any speech data was recorded. Unfortunately, for ACC, the short-term spectrum of this initial noise response, NOISE1, had more boost in the low-frequency end relative to the high-frequency end than did the short-term spectrum of NOISE2, the noise response extracted from the very beginning of an utterance file, before the

SYSTEM	FILTERS	RECOGNITION ACCURACY	
		95 DB	105 DB
ACC	400 Hz HPF, 1.5 kHz LPF, 3.2 kHz LPF	90.8%	--
VOUGHT	5 kHz LPF	92.8%	72.8%, 60.9% (Two Tests)
M12	5 KHz LPF	95.4%	90.3%
(ACC, M12) MODIFIED NOISE1 UNMODIFIED NOISE1	200 Hz HPF, 5 kHz LPF	-- --	39% 73.3%

Table 19. Recognition accuracies obtained for speaker RS using the Verbex 4000 on the 13-word minimal pairs vocabulary. "Modified Noise 1" and "Unmodified Noise 1" are explained in the text.

onset of speech. The reasons for this change in response are not clear, although it is possible that the shape of the background noise itself might have shifted during the recording session. This difference in shape between the short-term spectra for NOISE1 and NOISE2 for ACC caused a difference in shape between the spectra for the NOISE1 and NOISE2 that were created digitally for (ACC, M12). A modification in the level of ACC's contribution to NOISE1 for (ACC, M12) brought the shape of its short-term spectrum more in line with that of NOISE2. The test employing this modified NOISE1 file for background noise estimation yielded a recognition accuracy of 39%; on the other hand, when an (ACC, M12) NOISE1 file incorporating no such modification ("unmodified NOISE1") was used, the test yielded an accuracy of 73.3%. Another inconsistency is apparent in the recognition accuracies found for Vought. The two test runs in 105 dB, which were performed identically, yielded recognition accuracies that are 11.9% apart. Interestingly, M12 yielded good recognition accuracy even in 105 dB. The reasons for the discrepancies in performance noted above are unclear; therefore, because the Verbex unit's performance with 105 dB data was unreliable, no conclusive comparisons among single-sensor and two-sensor systems can be drawn.

8. FEATURE-BASED PARALLEL-INPUT MULTISENSOR SPEECH RECOGNITION

In the preceding sections, we described our work on single-input multisensor systems. In this and the following section, we present our work that demonstrates the feasibility of using a parallel-input multisensor system for high-performance speech recognition. From the results of our long-term and short-term spectral analyses of the various sensors presented in Section 2, we have demonstrated that different sensors provide different kinds of information about various speech sounds. In addition, some sensors are more immune to noise than others in a given frequency band. To test the feasibility of using multisensor information in a parallel-input speech recognition system, we investigated two different approaches. In the first, we explored how the strengths of the different sensors in transducing certain phonemes might be exploited in feature-based phonetic discrimination tests. If the use of features from multiple sensors proved successful, the features could be incorporated into a feature-based speech recognition system. In the second approach, called the long-vector approach, we obtained the parameter data for the parallel-input multisensor system from individual sensor parameter data by merging them on a frame-by-frame basis, thus generating "long vectors" of parameters. We then tested the single sensor data and the long-vector data using our hidden Markov model-based speech recognition research system.

For our parallel-input multisensor speech recognition work, we used as sensors the Vought, M12, a throat accelerometer (position 10), and a nasal accelerometer; as vocabulary, we used all or selected subsets of our minimal pairs database (see Section 6.2 and Appendix A). Below, we present the feature-based approach in this section and the long-vector approach in Section 9.

8.1 Acoustic-Phonetic Experiment Facility

In our investigation of the feature-based approach, we used BBN's Acoustic-Phonetic

Experiment Facility (APEF) [6]. The APEF program allows a researcher to formulate and perform an experiment in acoustic-phonetics in a short time. The capabilities built into APEF that make it particularly fast and convenient to use include the following:

1. A highly interactive, powerful English-like programming language allows the researcher to define the phonetic environments of interest and specify how to compute the features for a given experiment.
2. APEF can access a large database very quickly.
3. APEF performs various statistical analyses of features computed in an experiment. It also performs recognition tests and displays both the data and the results.
4. The resulting acoustic-phonetic features and recognition algorithms can be transferred directly to recognition programs.

8.1.1 Performing an APEF Experiment

If it is desired to find useful acoustic-phonetic features for a phonetic discrimination task, the following steps are carried out using APEF. The search for features to test is performed by plotting some precomputed parameters for several examples of the phonetic context of interest, for example, voiced fricatives followed by vowels. Where significant differences in the same parameter exist among the phonemes to be discriminated, a feature can be specified. Once a set of features to test has been decided upon, a procedure to calculate the features can be written and edited using the programming language included in APEF. Because APEF "understands" the programming language used, it tries to catch logical errors in the procedure. All or some logical subset of the database is then searched; for each occurrence of the phonetic context found, the procedure is run to compute the desired features. Statistical analyses can be performed on the feature data collected. The results of these analyses can be displayed in a variety of plots, tables, or graphs. Finally, "fair" recognition tests that use jack-knifing and a multivariate classifier are performed. Training involves the fitting of a Gaussian distribution to all the data points collected for each feature under study. During testing, each

sample is first removed from the multivariate distribution, then classified according to the modified distribution. From the test results, the best set of features for performing the phonetic discrimination can be determined.

8.1.2 Typical APEF Experiment

To provide a better understanding of how the APEF program was used to perform tests for the feature-based parallel-input system, we present here an example of a typical APEF experiment.

A listing of the algorithm used to perform the discrimination between initial voiced plosives and initial unvoiced plosives with Vought in 95 dB noise is given in Fig. 12. This experiment tried to exploit the fact that the puff noise energy in the Vought signal was larger for unvoiced plosives than for voiced plosives. The first line of the algorithm lists the phonetic context of interest, which in this example is voiced or unvoiced plosive followed by a vowel; the algorithm is only executed on those utterances having this context. Steps 1 through 12 specify the features and related quantities to be calculated for use in discrimination tests.

Fig. 13 gives two examples of the plots, created by steps 13 through 18 of the algorithm, for the utterances DODE and TOAD, respectively. LEZ is low-frequency energy; MEPZ is mid-frequency energy. Appendix B contains the definition of LEZ, MEPZ, and a number of other parameters and features we used. The differences in puff noise energy between the two utterances shown is apparent in the LEZ plots between the times "back 5" and "bot". Also shown at the top of Fig. 13 are the values calculated in each step of the algorithm for the two utterances.

Finally, Fig. 14 illustrates two APEF commands employed in the search for useful features. The "Statistics Table" command lists a feature's statistics grouped according to the phonetic classes being discriminated. The "Gather Distr for Discrimination" command performs a test to discriminate among the classes of interest. Included in the listing of test results are the definitions of all features used, the percentage of utterances correctly classified, and a confusion matrix for the classes being discriminated.

```
For Context  (voiced plosive unvoiced plosive) (vowel)
Tabulate
 1) Change Parameter File Extension to PR3
 2) tmax
   Time of maximum of parameter Derivative of MEPZ from 10 until 60
 3) bot
   Next time that parameter Derivative of MEPZ from tmax until 0
        is less than 0.5
 4) back 5
   Difference of bot and 5
 5) mstartavg
   Average of parameter MEPZ from back 5 until bot
 6) lstartavg
   Average of parameter LEZ from back 5 until bot
 7) mmin
   Minimum of 5 and back 5
 8) msilavg
   Average of parameter MEPZ from 0 until mmin
 9) lsilavg
   Average of parameter LEZ from 0 until mmin
10) mdiff
   Difference of mstartavg and msilavg
11) ldiff
   Difference of lstartavg and lsilavg
12) lvsmdiff
   Difference of ldiff and mdiff
13) Set up parameter plot between 0 and 100
14) Plot parameter LEZ with lower limit 10 and upper limit 80
        in Y-space of 200 with axis name LEZ
15) Plot parameter MEPZ with lower limit 10 and upper limit 80
        in Y-space of 200 with axis name MEPZ
16) Vertical Cursor on plot at tmax
17) Vertical Cursor on plot at bot
18) Vertical Cursor on plot at back 5
```

FIG. 12. APEF algorithm listing for the discrimination between voiced plosives and unvoiced plosives in the word-initial position.

```
D>list all results

 1)     50 T, 80 OW, 90 TOADS02,
  1) 0,  2) tmax: 33,  3) bot: 30,  4) back 5: 25,  5) mstartavg: 26.2448,
  6) lstartavg: 43.6602,  7) mmin: 5,  8) msilavg: 18.4765,
  9) lsilavg: 28.7521,  10) mdiff: 7.76836,  11) ldiff: 14.9081,
  12) lvsmdiff: 7.13973,  13) 0,  14) 0,  15) 0,  16) 0,  17) 0,  18) 0,

 2)     50 D, 80 OW, 90 DODES02,
  1) 0,  2) tmax: 32,  3) bot: 26,  4) back 5: 21,  5) mstartavg: 18.3932,
  6) lstartavg: 27.6371,  7) mmin: 5,  8) msilavg: 19.4214,
  9) lsilavg: 29.2167,  10) mdiff: -1.02819,  11) ldiff: -1.57963,
  12) lvsmdiff: -0.551441,  13) 0,  14) 0,  15) 0,  16) 0,  17) 0,  18) 0,
```

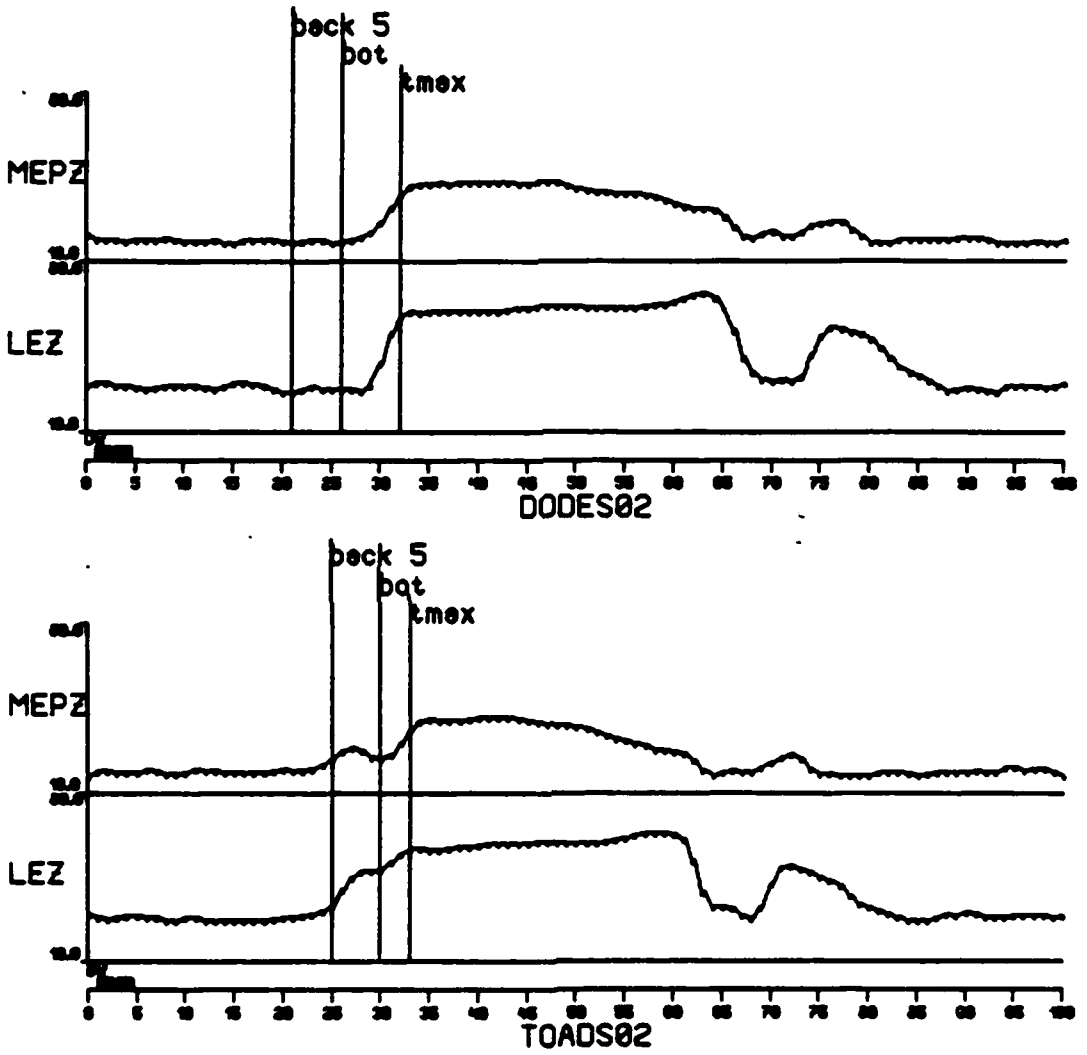


FIG. 13. Examples of APEF plots and algorithm results for two utterances.

```
D>statistics table
: lvsmdiff
 12) lvsmdiff:
   Difference of ldiff and mdiff
Label Determiner: functions # of classes: 2
13) Function: check label on segment: #: 1
with features: (voiced.plosive )
 13) Check if label on segment # 1 was (voiced.plosive) [confirm]
Use this one?yes
Use true values? yes
Another filter function? no
Use true results? yes
240 samples are in this class
Class name: vp
Remaining - 80 samples are in this class
Class name: up

Class   #    Avg    Std      Min     10     90     Max

vp      240  -8.2    1.6     -4.8   -2.2    1.9    4.6
up       80  11.8    4.9     -1.1    5.4   17.2   23.3

D>gather Distr for Discrimination

How many features: 1
Using features 1) : lvsmdiff

 12) lvsmdiff:
   Difference of ldiff and mdiff
Label Determiner: same as last time Also discriminate classes now? yes

 12) lvsmdiff: 1.000000

          vp   up
      |-----------
   vp |   238    3
   up |     2   77

D>gather Distr for Discrimination

How many features: 1
Using features 1) : ldiff

 11) ldiff:
   Difference of lstartavg and lsilavg
Label Determiner: same as last time Also discriminate classes now? yes

 11) ldiff: 1.000000

          vp   up
      |-----------
   vp |   237    3
   up |     3   77
```

FIG. 14. An APEF statistics table and two discrimination tests.

8.2 Data Preparation

Before we could use our APEF program to find features for the parallel-input system, a number of preparations had to be made. First, for speaker RS's 95 dB minimal pair data, we created parameter files with our utility program called PSA. This program calculates 44 parameters per 10 ms frame of the waveform. Included among these parameters are energy measures such as R0 as well as first, second, and third formant frequencies. One parameter file for each utterance/sensor combination was created. Second, we generated a label file for every utterance. A label file contains the phonetic spelling of the speech contained in the corresponding waveform file; because each APEF experiment operates only on data occurring within a user-specified phonetic context, the program needs the phonetic spelling of each utterance to determine which utterances should be included in a given experiment. Third, in order that we could use features derived from more than one sensor signal in recognition experiments, we modified the APEF program to enable it to process multiple parameter files for each utterance.

8.3 Phonetic Discriminations We Tested

Our selection of minimal pair distinctions to study was influenced by two factors: 1) which distinctions gave trouble for one or more sensors in the Verbex minimal pairs tests, and 2) which distinctions we expected would be helped by the use of multisensor information. Table 20 gives the phonetic discriminations we chose to use for testing the feature-based parallel-input system, along with abbreviations to refer to the discriminations conveniently. All initial-position phonetic discriminations we studied involved the two vowel contexts, [o] (as in MODE) and [ɛ] (as in BET); for final-position discriminations, B-D-G/F involved the vowel [a] (as in SOB) only and NAS-VP/F involved both [a] and [o]. We note that in all cases, discriminations were performed without regard to the vowel. To see what minimal pair words we used in various discriminations, consult Table 4.

DISCRIMINATION	ABBREVIATION	# OF TOKENS PER CLASS
[m] vs. [n] (initial position)	M-N/I	M: 38 N: 40
[b] vs. [d] vs. [g]		
(initial position)	B-D-G/I	B: 60 D: 140 G: 40
(final position)	B-D-G/F	B: 20 D: 20 G: 20
nasals vs. voiced plosives		
(initial position)	NAS-VP/I	NAS: 78 VP: 240
(final position)	NAS-VP/F	NAS: 99 VP: 420
voiced vs. unvoiced plosives (initial position)	VP-UVP/I	VP: 240 UVP: 80
[p] vs. [t] vs. [k] (initial position)	P-T-K/I	P: 40 T: 20 K: 20

Table 20. Selected phonetic discriminations included in our study.

8.4 Selection of Features

We did not place any constraints on the choice of features that could be used in a given discrimination; whatever features yielded the best performance for each sensor or sensor combination were chosen, regardless of the features used for the same discrimination with other sensors. The "best" feature combinations were usually found with APEF's "Find Optimal Features" option. In this option, APEF executes the "Gather Distr for Discrimination" command for all possible combinations of features derived from a given feature set; the minimum and maximum number of features allowed in the combinations are set by the user. APEF then prints out the results for the feature combinations from best to worst based on discrimination performance. Because of time constraints, not all possible sensors or sensor combinations were tried for every discrimination. Also, the set of "optimal features" we found for each discrimination was based only on those features that we could find

and test within a few hours; with more study, additional useful features might have been found. Finally, for many of the discriminations listed, more than one set of features gave results equal to or nearly equal to those shown. For example, for VP-UVP/I with throat accelerometer, two features together ("lezcdiff" and "mrise") yielded a 96.9% result, which was only 0.3% less than the result obtained with the single feature "lezclose". For a description of features we used, refer to Appendix B.

For some of the M-N/I and B-D-G/I discriminations studied, we found that large, abrupt changes (hereafter referred to as "glitches") in smoothed formant frequencies F2M and F3M (see Appendix B) frequently occurred where formant features were being measured. This glitching was noticeable for the throat accelerometer as well as for the two microphones in 105 dB. For those discriminations where glitching was observed for the sensors of interest, modifications were made to the formant feature measurements to exclude glitches. We also conducted experiments where no glitch-detection was performed. In some cases, employing glitch-detection improved performance; in others, the opposite was true. In a real system, the formant tracker could be optimized for each sensor individually to minimize glitching.

For some phonetic discriminations studied, we could find "no useful features" for a particular sensor. For our purposes, we define "no useful features" to mean either that 1) our visual inspection of the parameters available in APEF showed no significant differences among the phonemes to be discriminated or that 2) testing of features that looked promising never yielded performance better than 70%, either alone or in combination with other features.

The primary goal of our APEF investigation was to demonstrate that using multiple sensor signals leads to higher recognition accuracies in selected phonetic discrimination tests than using single sensor signals does, under high-noise conditions. Within this goal, we developed a set of reasonable features listed in Appendix B. For each phonetic discrimination and for each sensor or sensor combination, we determined the optimal features combination as described above. If the performance of a certain optimal features combination was not satisfactory, it is quite possible that by developing additional features through further work, one may have been able to improve the performance to an acceptable level. We did not,

however, attempt to do this because of the limited scope of the project. It is in this sense that our investigation is not complete or exhaustive. However, we believe that the test results presented below for 105 dB noise show that we achieved our primary goal stated above.

8.5 Test Results in 95 dB Noise

The results of the phonetic discrimination experiments performed with APEF for the 95 dB data are listed in Table 21. (We did not include M12 in our tests in 95 dB noise.) For a given discrimination, we list for each sensor only the highest performance score achieved for all features and feature combinations tested. Examination of Table 21 clearly shows that Vought did very well by itself in 95 dB noise for all of the phonetic discriminations we studied. In no case did the throat accelerometer yield better performance; the only discriminations for which another sensor performed better than Vought was NAS-VP/F. Here, the nasal accelerometer was superior. Overall, these results support our previous finding obtained with the Verbex 4000 that 95 dB noise is not a problem for the Vought microphone. Table 22 lists the features used in the 95 dB tests shown in Table 21. Descriptions of these features can be found in Appendix

Since Vought alone performed well in 95 dB for all discriminations studied, the value of combining features from multiple sensors was unclear. Therefore, to make the discriminations more difficult, we chose to increase the background noise level for both microphones to 105 dB, using the procedure described in Section 7.4.2. Because the accelerometers are essentially insensitive to acoustic background noise, we used the same parameter files for the accelerometers for both the 95 and 105 dB tests.

8.6 Tests in 105 dB Noise

Table 23 gives recognition accuracies for single and multiple sensors in 105 dB noise. We have used the notation {V,M} to denote the two-sensor parallel-input system involving Vought

TEST	SINGLE SENSORS			BEST SINGLE SENSOR
	V	A	NA	
M-N/I	97.4%	94.9%	**	V 97.4%
B-D-G/I	100%	90.4%	*	V 100%
B-D-G/F	98.3%	96.7%	*	V 98.3%
NAS-VP/I	99.1%	**	99.1%	V or NA 99.1%
NAS-VP/F	96.3%	*	98.1%	NA 98.1%
VP-UVP/I	98.4%	97.2%	*	V 98.4%
P-T-K/I	97.5%	**	*	V 97.5%

Table 21. Performance of single sensors in selected phonetic discrimination tests listed in Table 20, in 95 dB noise. The abbreviations V, A, and NA denote, respectively, the Vought microphone, the throat accelerometer, and the nasal accelerometer. The symbol * indicates cases that were not investigated and the symbol ** indicates cases that did not have any useful set of features.

TEST	V	A	NA
M-N/I	f2diff f3diff	f2diff f2vow f3diff f3vow cdiff	**
B-D-G/I	f2diff f2vow f3diff f3vow cdiff	f2diff f2vow f3diff f3vow cdiff consilmepz	*
B-D-G/F	f2diff f2vow f3diff f3vow	f2trans f3diff f3trans cvow	*
NAS-VP/I	consnt energy	**	consnt energy
NAS-VP/F	consnt energy diff	*	nasaccdiff
VP-UVP/I	lvsmdiff	lezclose	*
P-T-K/I	zcdiff lezdiff vot mepzdiff	**	*

Table 22. Best feature sets for single sensors in 95 dB noise. (See Appendix B for definitions of the features listed.)

and M12. (In contrast, our notation (V, M) denotes the two-sensor system that mixes the Vought and M12 signals to yield a single speech input.) Table 24 lists the features used in the discriminations tests shown in Table 23.

For M-N/I and B-D-G/I, a combination of features from two sensors was found to give the best results, while in all other categories (except for P-T-K/I) one of the accelerometers gave the best performance. Compared to the results in Table 21 for 95 dB noise, the results in Table 23 for 105 dB noise show a degradation in Vought's performance of 2.5% or more in every category. (It is interesting to note that, although our short-term spectral analysis of the Vought signal indicated that it does not transduce nasals and voiced plosives well, Vought's APEF tests of M-N and B-D-G at both noise levels yielded a performance of 88.3% or better.) For P-T-K/I, the features that worked well for Vought in 95 dB could not provide better than 62% performance for Vought in 105 dB. Features tried for M12 gave roughly the same performance. Because of these poor results, the {V,M} combination was not tried for P-T-K/I in 105 dB.

In all categories shown in Table 23, the "best overall" case yielded fewer than one-half the number of errors made by either microphone alone. The Vought microphone alone performed reasonably well for all discriminations, yielding accuracies ranging from 88.1% to 94.9%. The reason for this high performance is that the features used were chosen separately for each phonetic discrimination, as we mentioned above. However, the "best overall" case reduced recognition errors to between one-half and one-twelfth of the number produced by Vought; this substantial improvement is due to the additional freedom to select the best sensor or sensor system for each phonetic discrimination. Furthermore, the accuracy produced by the "best overall" case is excellent with a range from 96.7% to 99.1%, at 105 dB noise. We may interpret the "best overall" case as one in which we select the sensors and features to use as a function of the phonetic context in question; this may be possible in a two-pass recognition algorithm, with a conventional single microphone system performing the first pass and the parallel-input system performing the second pass. These results indicate that a feature-based recognition system incorporating features from all four sensors is likely to perform substantially better than a system using features from one sensor only. We hasten to add,

TEST	SINGLE SENSORS				BEST SINGLE SENSOR	MULTIPLE SENSORS		BEST OVERALL
	V	M	A	NA		{V, M}	{A, V}	
M-N/I	94.9%	*	94.9%	**	A or V 94.9%	*	98.7%	{A, V} 98.7%
B-D-G/I	88.3%	94.2%	90.4%	*	M 94.2%	96.7%	94.6%	{V, M} 96.7%
B-D-G/F	93.3%	90.0%	96.7%	*	A 96.7%	***	*	A 96.7%
NAS-VP/I	89%	89.6%	**	99.1%	NA 99.1%	96.5%	***	NA 99.1%
NAS-VP/F	88.1%	82.7%	*	98.1%	NA 98.1%	*	***	NA 98.1%
VP-UVP/I	94.4%	**	97.2%	*	A 97.2%	***	***	A 97.2%

Table 23. Performance of single and multiple sensors in selected phonetic discrimination tests listed in Table 20, in 105 dB noise. The abbreviation M denotes the microphone M12, and the symbol * indicates cases in which using two sensors yielded no improvement over either single sensor. For other notations used, see the caption of Table 21.**

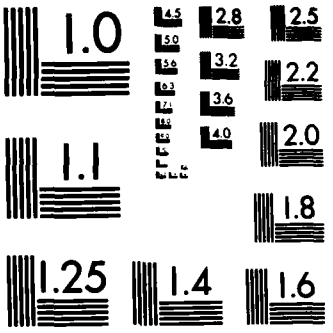
TEST	V	M	{V, M}	{A, V}
M-N/I	f2diff f2vow f3diff f3vow	*	*	A: f2diff f2vow V: cdiff f3diff f3vow
B-D-G/I	f2diff f2vow f3diff f3vow cdiff	f2diff f2trans f3diff f3trans	V: f2diff cdiff M: f2diff f2trans f3diff f3trans	A: f2diff f2vow V: f2diff f2vow f3diff cdiff
B-D-G/F	f2vow f3vow cdiff	f2diff f2vow f3vow	***	*
NAS-VP/I	consnt energy slope	diff slope	V: consnt energy slope M: consnt energy slope	***
NAS-VP/F	burstdiff diff	burstdiff diff	*	***
VP-UVP/I	ldiff	**	***	***

Table 24. Best feature sets for single and multiple sensors in 105 dB noise. (See Appendix B for definitions of the features listed.)

however, that for a complete feature-based isolated-word recognition system, the recognition accuracy is likely to be less than the highest figures reported above for individual discrimination tests.

AD-A174 693 NOISE-IMMUNE MULTISENSOR TRANSDUCTION OF SPEECH(U) BBN 2/2
LABS INC CAMBRIDGE MA V R VISWANATHAN ET AL AUG 86
BBN-6114 RADC-TR-86-87 F30602-84-C-0088

UNCLASSIFIED F/G 9/1 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

9. LONG-VECTOR APPROACH TO PARALLEL-INPUT MULTISENSOR SPEECH RECOGNITION

The long-vector approach presented in this section is a simple way of taking advantage of multiple, parallel inputs; also, it allows the use of an existing speech recognition algorithm for testing its effectiveness. This approach consists of forming, on a frame-by-frame basis, a composite or long vector of parameters by simply collecting together the parameters extracted from each of the parallel inputs and evaluating the long-vector data using an existing speech recognition algorithm. In our investigation of the long-vector approach, we used BBN's research speech recognition system, which uses vector quantization and a discrete hidden Markov model. Below, we describe the method we used for extracting parameters from individual sensor signals, review our vector quantization algorithm, describe BBN's hidden Markov model-based recognition system, and present the results of recognition tests of parallel-input multisensor systems, individual sensors, and a single-input two-sensor system.

9.1 Parameter Extraction

Because we wished to select information to be included in the long vectors based on knowledge about the frequency location of this information for each sensor, we chose to use parameters based in the frequency domain. Then, for example, if the information in certain frequency bands for the accelerometer were buried in noise, as we would have expected in the higher frequencies, we could include only the spectral-band parameters for the lower frequencies in the long-vector data; in this way, we would eliminate the noisy bands from further consideration. Generalizing this procedure, we could assign different weights to different spectral bands, reflecting, for example, our prior knowledge of the quality of information in individual spectral bands. We used Mel-frequency warped and cepstrally smoothed spectral-band parameters, since Mel-frequency warping has been found to yield good speech recognition performance [7]; cepstral smoothing was performed to remove the pitch structure from the short-term spectra.

The procedure we used to compute the spectral-band parameters is as follows. For each 10 ms frame of the waveform file, we calculate the cepstrum of the signal over an analysis interval of 20 ms by

1. calculating the log magnitude spectrum $P(w)=\log|X_n(w)|$;
2. performing energy normalization on the spectrum using the formula $P'(w)=P(w)-\langle P(w)\rangle$, where $\langle P(w)\rangle$ is the average over frequency;
3. Mel-warping the frequency scale of the spectrum, using the transformation $f_{mel} = \log_2(1 + f_{Hz}/1000)$; and
4. taking the inverse discrete Fourier transform (DFT).

Energy normalization is performed so that the cepstral values obtained are independent of the signal energy. After retaining the first 15 points of the cepstrum and setting the rest of the points to zero, we calculate the DFT of the truncated cepstrum. The average power values, in dB, of the DFT points in 25 equally spaced bands on the Mel-warped frequency scale yield the spectral-band parameter values of interest. These bands cover the frequency range 0-5000 Hz. Because of the Mel-warping, the spectral bands below 1 kHz are narrower than those above 1 kHz; in this way, the set of spectral-band parameters place greater emphasis on the signal's low-frequency information than on its high-frequency information. Table 25 lists the center frequency in Hz for each of the 25 spectral bands. We note that we lowpass filtered each sensor signal at 5 kHz and digitized at 10 kHz; no other filtering was applied to the digitized waveform files before calculating the spectral-band parameters.

Those spectral-band parameters that are considered most useful for each sensor are chosen for inclusion in the long-vector parameter files. For each 10 ms frame, the parameters of interest from each sensor are concatenated to form a single long vector; the resulting long vectors are written into a new parameter file. We note that if testing of a single sensor is desired, the creation of long-vector parameter files is not necessary.

BAND	CENTER FREQUENCY, HZ
1	36
2	114
3	196
4	285
5	381
6	483
7	593
8	712
9	839
10	976
11	1122
12	1280
13	1450
14	1632
15	1827
16	2037
17	2263
18	2505
19	2766
20	3045
21	3346
22	3669
23	4016
24	4389
25	4789

Table 25. Center frequencies of the 25 spectral bands we used.

9.2 Vector Quantization

The parameter vector representing a frame of speech is quantized to the nearest vector from a codebook of templates; usually we use codebooks of 256 templates. This vector quantization process is necessary because our speech recognition system uses discrete hidden Markov models, which model the statistical structure of discrete random processes. We explain in the following section the details of hidden Markov models.

The codebook is obtained by using a clustering algorithm on a "training set" of parameter vectors to determine the required number of templates. We typically use 5,000 to 100,000 vectors to determine codebooks with 256 to 1024 templates. In this research effort, we used a hierarchical clustering algorithm that defines a binary tree on the templates to speed up the quantization of a vector (instead of computing n distances to determine the nearest template, only 2 $\log_2$ n distances are required with the tree). The binary clustering algorithm requires two orders of magnitude less computation than the traditional full-search K-means clustering algorithm [8]. The computational savings is particularly important in our case because each choice of the long vector requires the design of a new quantizer. We expect the loss in performance due to our use of the binary tree instead of the full search to be small. The distance measure we used in vector quantization was the weighted Euclidean distance, although we used unity weights in our investigation.

9.3 Discrete Hidden Markov Model-Based Speech Recognition

9.3.1 Hidden Markov Model

In evaluating the long-vector approach for parallel-input multisensor speech recognition, we used a speech recognition research system that was developed in an earlier IR&D project. The research system is flexible and can perform both isolated-word recognition and connected-word recognition. It also supports the use of a syntax for a particular task. The system uses a discrete hidden Markov model for each word in the vocabulary.

A discrete hidden Markov model (HMM) is described by the following items:

- A set of states, called S, in which we distinguish two particular states: i) 0, the initial state, and ii) LAST, the final state.
- A transition matrix A over S-{LAST} x S-{0}, where the element A(i,j) is the a priori probability of a transition from state i at time t to state j at time t+1.
- An output matrix B, where each column is a discrete probability density function (PDF) associated with each state of S-{0}. B(k,j) is the probability of observing output symbol k at time t if the state at time t is j.

The model starts in the initial state. Then, at each time instant t, it selects the next state based on the previous state and the transition matrix. Assume that the model goes to state j at time t given that it was in state i at time t-1 (the probability of selecting j is A(i,j)). At state j, the model will select an output symbol according to the density B(k,j); the output symbol $y(t)$ is the observed quantized spectrum at time t.

When a sequence of output symbols $y(1), y(2), \ldots, y(T)$ (quantized spectra) is observed, we need to determine the probability that a given hidden Markov model produced that sequence. Note that many state sequences $x(1), x(2), \ldots, x(T)$ can produce the same observed sequence. Therefore, the probability of the observed sequence is

$$Pr[y(1), y(2), \ldots, y(T)] = \Sigma\, Pr[y(1), y(2), \ldots, y(T) \mid x(1), x(2), \ldots, x(T)],$$

where the summation is over all values of $x(1), x(2), \ldots, x(T)$. To compute these probabilities efficiently we use the forward-backward algorithm [9, 10]. To estimate the model parameters from a set of tokens for a word, an iterative algorithm is used to adjust the transition matrix and the output matrix to increase the likelihood of observing the training tokens for that word, given the model at each iteration.

To recognize a given sequence of observations, for every word model we compute, using the Viterbi algorithm [10], the state sequence most likely to produce that sequence of observations. The recognition algorithm selects the word model that has the largest "a-posteriori" state sequence probability (i.e., given the observations) as the recognition decision.

In previous work performed with our research system, we found that allowing 3 states per phoneme gave good recognition performance. Because each minimal-pair word consists of three phonemes, we chose to employ a 9-state model for the words. We used the 9-state model for the pause found before and after each word as well; no attempt was made to optimize the number of states used in the model.

9.3.2 Training of the HMM Recognizer

Training the HMM recognizer involves computing the values of the state transition probabilities (9 x 9) and output symbol probabilities (9 x N) for each frame of input, where N is the number of output symbols; N=256 for 8-bit vector quantization. The input files ("tokens") contain the output symbol for each frame of the utterance as chosen by the vector quantizer. To train a word model, we do not use endpoint detection. Instead, we use a pause model before and after the word model and train all three models simultaneously. The pause model is trained initially using 15 frames of silence at the beginning of two tokens for each vocabulary word (a total of 60 tokens). Each vocabulary word model, along with the pause model, is then trained on 10 tokens. Ten iterations of training were performed on the entire set of 10 training tokens to obtain a word model.

9.3.3 Testing with the HMM Recognizer

Testing consists of finding for each test token which vocabulary word (including optional initial and final pauses) was likely to have caused the test token's sequence of quantized output symbols. Testing was performed on the 10 tokens per vocabulary word that were not used for training.

9.3.4 Performance of Our HMM Speech Recognition Research System

Our non-real-time HMM speech recognition research system, was developed at BBN in an IR&D project. In tests of the system performed in that project, we found that 1) the system yielded a 99% recognition accuracy for speaker-dependent recognition of the E set, ahd 2) for speaker-independent recognition, the system had digit recognition rates of 98.6% for isolated digits and 95.5% for connected 7-digit strings. These test results clearly show that our research system performs speech recognition very successfully.

9.4 Selection of the Vocabulary

Because a large vocabulary increases the difficulty of the recognition task with single sensors, it provides more room for possible improvement in recognition performance with multiple sensors. Therefore, a large vocabulary was desirable for the long-vector approach, and we could handle a large vocabulary because our research system places no constraints on vocabulary size, unlike the Verbex unit. However, we also wished to minimize what we hypothesized was a durational problem with the minimal pairs vocabulary, which was first encountered during the Verbex tests (see Section 7.3.5). By truncating (or "chopping") the parameter files for each sensor at a frame located partway into the vowel of each utterance, we sought to both minimize the durational problem and retain a large vocabulary size.

Because we wished to use "chopped" data, it was necessary to eliminate from the original 44-word minimal pairs vocabulary those words whose initial consonant and vowel duplicated

those of another word; for example, because the original vocabulary contained both "sod" and "sog", the word "sog" was removed. The resulting 30-word vocabulary is listed in Table 26.

Met,	Heed,	Doze,	Leaf,	Sod,	Hid,	Goad,
Sin,	Code,	Bet,	Debt,	Get,	Hood,	Let,
Head,	Mode,	Shod,	Ret,	Net,	Load,	Wet,
Pode,	Yet,	Fin,	Had,	Toad,	Pet,	Hud,
Hod,	Node.					

Table 26. A 30-word minimal pairs vocabulary.

As discussed in Section 7.3.4, acoustic background noise at 95 dB was found not to be a problem for Vought and M12 for speaker RS's data. Therefore, for the purpose of our long-vector work, we chose to raise the noise level of RS's microphone data to 105 dB digitally, following the same procedure used to generate the 105 dB waveform files required for both our APEF work and our tests with the Verbex 4000. This procedure is outlined in Section 7.4.2. However, when we compared a plot of the R0 (energy) contour for CH for a typical utterance in 95 dB with the R0 contour for RS for the same utterance in 105 dB, both as transduced by M12, we found that the peak SNR for CH in 95 dB was very close to RS's peak SNR in 105 dB, and was about 13 dB. These plots are shown in Fig. 15. The similarities in SNR's can be attributed to a difference in speaking level between the two speakers as well as differences in microphone placement. Therefore, because the peak SNR's of RS in 105 dB and CH in 95 dB were comparable, we decided to use CH's 95 dB data without adjustments to its ambient noise level.

For both speakers, we used the throat accelerometer (ACC) and nasal accelerometer (NAS) data recorded in 95 dB, since the accelerometer is essentially insensitive to acoustic noise.

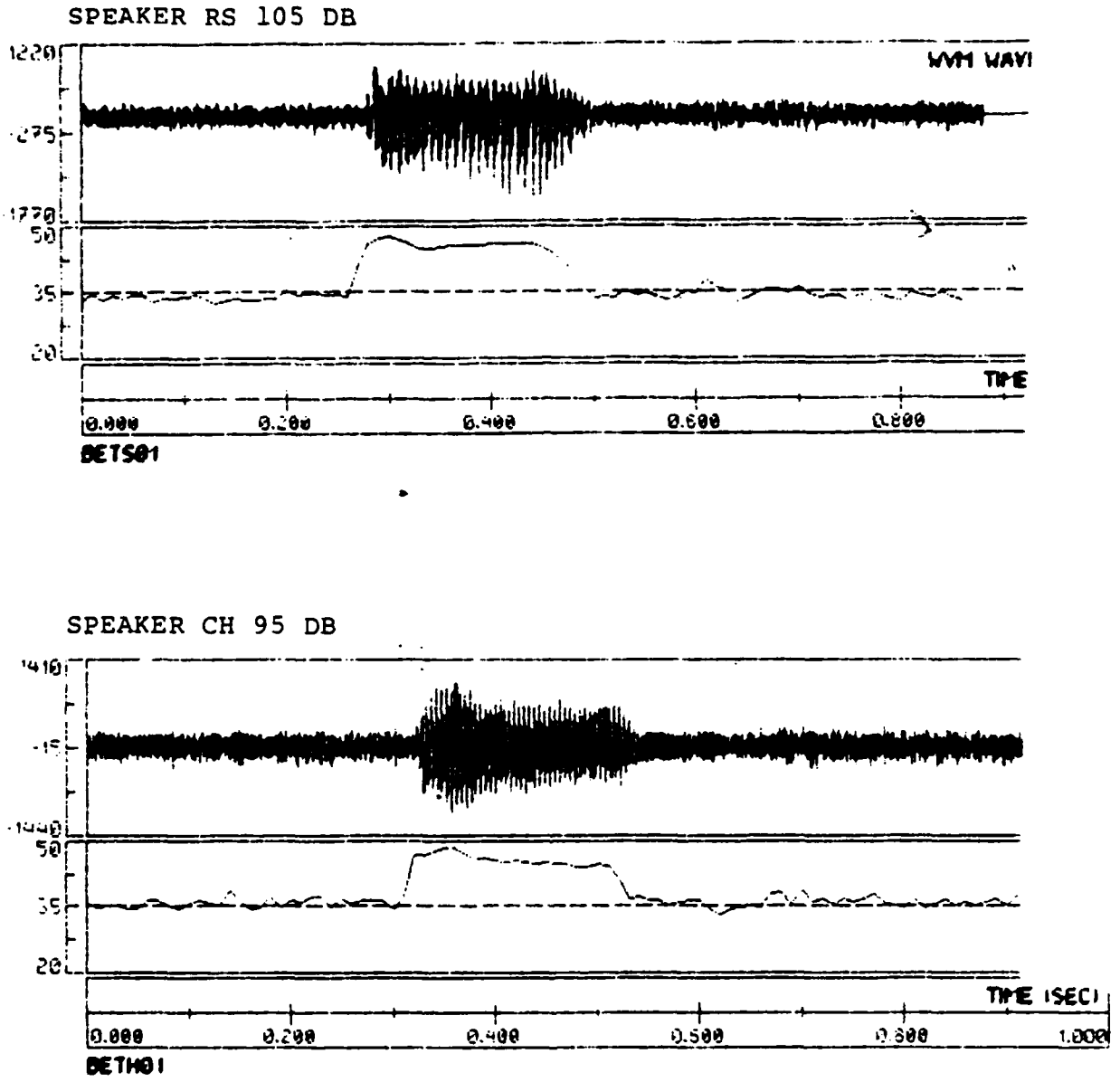


FIG. 15. Plots of waveforms and energy contours for the word "BET" transduced by M12 for speaker CH in 95 dB and speaker RS in 105 dB.

9.5 Recognition Tests

For convenience, we use in the text and tables a shorthand notation to describe single sensors and long-vector multisensor systems under test. We present two examples to explain this notation. [V(1-25)] indicates that for the Vought microphone, the spectral bands 1 through 25 were included in the single-sensor test being discussed. [ACC(1-15), M12(1-25)] indicates that the first 15 spectral bands for the throat accelerometer and all 25 bands for M12 were included in the long-vector system under test. All references to single-input multisensor systems, such as those used in tests with the Verbex unit, are enclosed in parentheses; i.e., (ACC, M12).

In addition, we note that many sensor systems were tested twice. If a second test is listed for a given system, it was performed by switching the training and test tokens (10 each per vocabulary word) used in the first test. Because the data for the two tests is non-overlapping, we may average the two test results together; then, we can consider the average as the result of a test of 20 tokens per word.

9.5.1 Tests on "Chopped" Data

Tests on "chopped" data were performed for speaker RS only. The "chopping" was performed by truncating the single-sensor parameter files before vector quantization was performed. This truncation was performed automatically, using a decision rule which found the beginning of the vowel in each utterance by searching backwards from within the vowel for the first frame where the derivative of R0 (energy) exceeded 4 dB. If no derivative greater than 4 dB was found, that frame having the largest derivative was chosen. Twelve was added to the index of this frame to find the index of the "chopping" frame; the four SS parameter files of interest for the given utterance (ACC and NAS in 95 dB, Vought and M12 in 105 dB) were then truncated at this "chopping frame". Vought's 95 dB parameter files were used to find the "chopping frame", because the truncation algorithm worked more reliably on 95 dB data than on 105 dB data.

The recognition accuracy obtained for the chopped single-sensor parameter data, which included all 25 spectral-band parameters, is 52% for M12, 50% for Vought, 78.5% for ACC, and 58.8% for NAS. It is therefore clear that for the 30-word chopped vocabulary, the recognition performance of all single sensors, except the ACC, is poor. We then performed one long-vector test for the chopped data. For this test, the long vectors contained the first 15 spectral-band parameters for ACC and all 25 spectral-band parameters for M12. The resulting recognition performance was 83.9%, which was significantly better than the performance of either ACC or M12 alone for the chopped files.

To investigate our hypothesis that "chopped" data should yield superior performance to "unchopped" data, we repeated the same 30-word test conducted for [M12(1-25)]'s chopped data on its unchopped data; a recognition accuracy of 87.2%, an improvement of 35.2% over the recognition accuracy found for the test of the chopped data, was achieved. This result refutes our hypothesis concerning the usefulness of chopped data; therefore, we conducted all our subsequent long-vector work with unchopped data only.

9.5.2 Tests on "Unchopped" Data

Tables 27 and 28 list the recognition accuracies for single sensors and long-vector configurations for speakers RS and CH, respectively. We can make several observations about the recognition accuracies we obtained. First, for speaker RS, [V(1-25)]'s recognition accuracy was roughly 8% poorer than [M12(1-25)]'s for the same training and test sequence ("Test 1" in Table 27). Comparison of plots of R0 in 105 dB for both microphones showed that Vought's peak SNR was roughly 3 dB less than M12's peak SNR; the difference in recognition performance we found can be attributed to this difference in SNR's. Because of its poorer performance and to reduce the number of different multisensor configurations to test, we chose not to include the Vought in any long-vector tests. Second, the recognition accuracies obtained for the data for the two speakers were roughly the same (within 2-3%) for the same configurations. In particular, using the single sensor [ACC(1-15)] led to a recognition rate of 87.8% for RS and 85.8% for CH; when the single sensor [M12(1-25)] was tested, the

SENSOR CONFIGURATION	RECOGNITION ACCURACIES		
	TEST 1	TEST 2	AVERAGE
[ACC (1-15)]	88.3%	87.3%	87.8%
[ACC(1-25)]	90.9%	89.3%	90.1%
[M12(1-25)]	87.2%	82.7%	85.0%
[V (1-25)]	79.5%	-	-
[NAS (1-25)]	82.9%	-	-
[ACC (1-15), M12 (1-25)]	94.3%	90.0%	92.2%
[ACC (1-25), NAS (1-25), M12 (1-25)]	94.6%	-	-

Table 27. Recognition accuracies obtained for the 30-word minimal-pairs vocabulary tested for speaker RS in simulated 105 dB ambient noise. Explanations of the notations used for sensor configurations and the differences between "Test 1" and "Test 2" can be found in the text.

SENSOR CONFIGURATION	RECOGNITION ACCURACIES		
	TEST 1	TEST 2	AVERAGE
8-BIT CODEBOOK			
[ACC (1-15)]	86.3%	85.3%	85.8%
[ACC (1-25)]	92.6%	-	-
[M12 (1-25)]	84.9%	88.3%	86.6%
[ACC (1-15), M12 (1-25)]	95.0%	92.0%	93.5%
[ACC (1-25), M12 (1-25)]	95.0%	-	-
10-BIT CODEBOOK			
[ACC (1-15)]	83.9%	-	-
[M12 (1-25)]	82.6%	-	-
[ACC (1-15), M12 (1-25)]	92.0%	-	-

Table 28. Recognition accuracies obtained for the 30-word minimal-pairs vocabulary tested for speaker CH in 95 dB ambient noise. Explanations of the notations used for sensor configurations and the differences between "Test 1" and "Test 2" can be found in the text.

recognition rate was 85% for RS and 86.6% for CH. Third, for both speakers, when ACC and M12 were combined using the long-vector method to form [ACC(1-15), M12(1-25)], the resulting recognition accuracy was substantially better than that of either sensor tested alone; the recognition rate increased to 92.2% for RS and 93.5% for CH. These results clearly show that the use of spectral-band parameters from multiple sensors in the long-vector approach for HMM recognition yields performance superior to that obtained when only the parameters from one of the constituent single sensors are used. In these experiments, the long-vector approach yielded a 40% lower error rate than that of the best constituent sensor for the multisensor system [ACC(1-15), M12(1-25)].

Fourth, comparisons between the recognition accuracies for [ACC(1-15)] and [ACC(1-25)] for both speakers show that including the parameters from the upper 10 frequency bands improved recognition performance. However, the 95.0% recognition accuracy obtained for CH for [ACC(1-25), M12(1-25)] is the same as that found for "Test 1" of [ACC(1-15), M12(1-25)], which used the same sequence of training and test tokens as did the [ACC(1-25), M12(1-25)] test. In other words, no improvement was realized by including those same 10 bands for ACC in the long vectors. In a similar example, even though [NAS(1-25)] achieved a recognition accuracy of 82.9% by itself, the inclusion of NAS's 25 parameters and bands 16-25 of ACC in [ACC(1-25), NAS(1-25), M12(1-25)] for speaker RS improved performance over [ACC(1-15), M12(1-25)] only marginally. We note that in the first example, by including 10 more parameters from ACC in the long vectors, we increased the number of parameters in each vector from 40 to 50; in the second example, by including the additional ACC and NAS parameters, we increased the vector length from 40 to 75 parameters. We believe that using more training tokens and a larger codebook size would enable the presence of these additional parameters in the long vectors to improve recognition performance. It is probable that the 50- and 75-parameter vectors are too large for an 8-bit codebook to quantize effectively.

9.5.3 Additional Tests

9.5.3.1 Spectral Resolution in Vector Quantization

Because [ACC(1-15)] and [M12(1-25)] were each quantized with an 8-bit codebook, the quantization of the long vectors in [ACC(1-15), M12(1-25)], which contains all 40 of the parameters, reduced the effective number of bits used for the quantization of each parameter. This loss of spectral parameter resolution caused an increase in the mean-squared quantization error for each parameter in the long vector. For speaker CH, Fig. 16 gives a plot of the total quantization error for [M12(1-25)] (curve 1), [ACC(1-15)] (curve 2), [ACC(1-15), M12(1-25)] (curve 5), and the individual contributions of ACC and M12 to the foregoing long vector (curves 3 and 4). By comparing curves (1) and (3), we find that when an 8-bit codebook was used for quantizing the long vector, we were effectively using a codebook size of 5 bits for the M12 component and 7 bits for the ACC component; in other words, M12 lost 3 bits of resolution and ACC lost 1 bit of resolution when they were combined into the long vector and quantized, compared to when each was quantized separately. The relative bit allocation between M12 and ACC in the long-vector approach can be controlled by adjusting the parameter weights. We note that this particular choice of bit allocation may be slightly under-representing the information in M12.

In an attempt to improve the spectral resolution for each sensor in [ACC(1-15), M12(1-25)], we increased the codebook size from 8 bits (256 templates) to 10 bits (1024 templates); we then clustered, quantized, and tested the data. We see from Fig. 16 that a 10-bit codebook for the long vector yields effective codebook sizes of 8.5 bits for ACC and 6.6 bits for M12; M12's effective codebook size is still slightly smaller than the 8-bit codebook size of the single-sensor cases. The recognition accuracy we obtained for CH with the 10-bit codebook was poorer by 3% than that obtained from the data quantized with an 8-bit codebook, as seen from Table 28. Because only 10 tokens per vocabulary word were used for training, we believed that the number of spectral templates, which was increased by a factor of four, was too large in this case to allow adequate training of each word model; this insufficient training led to degraded recognition performance. To confirm this reasoning, we also tested the single

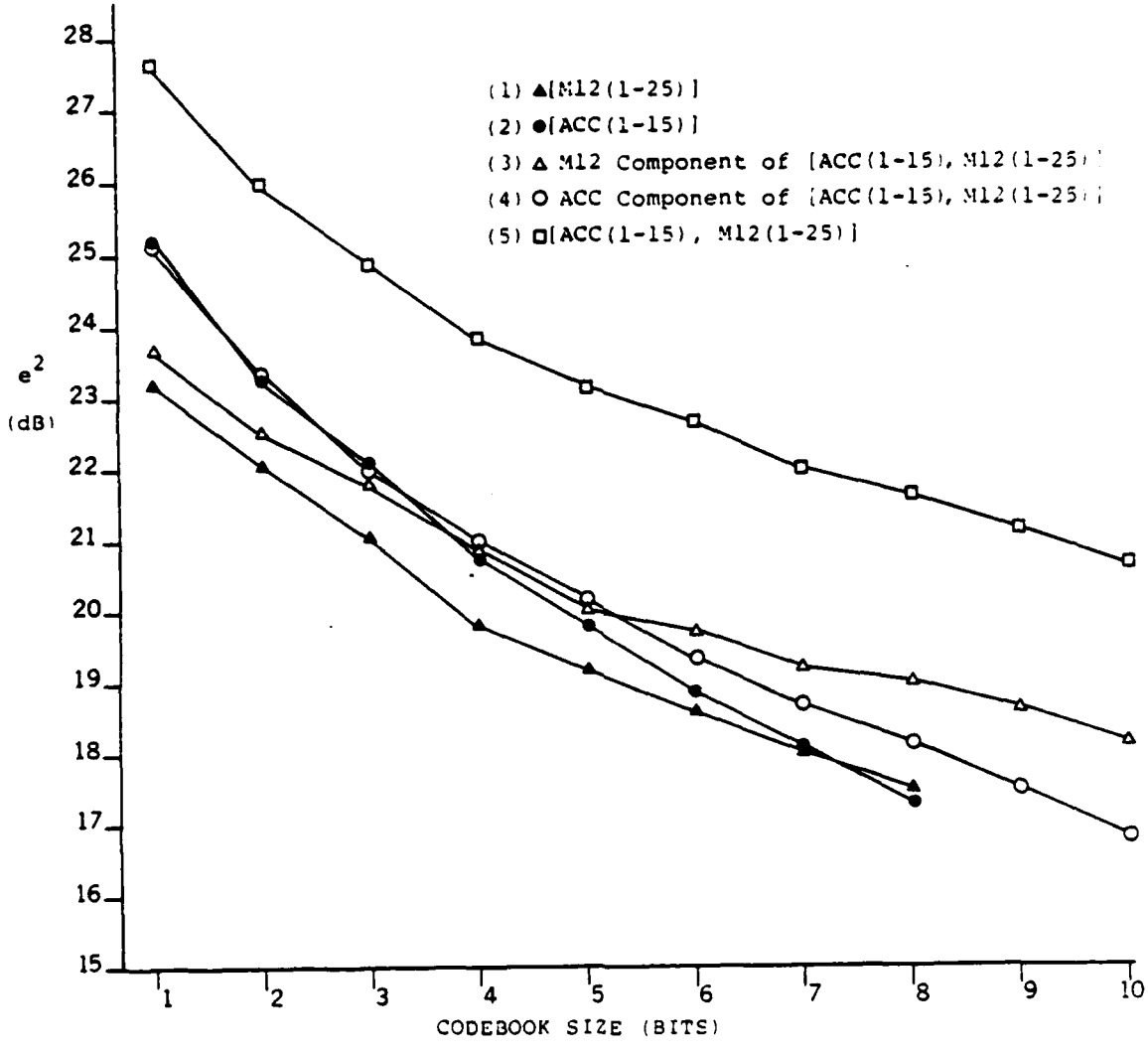


FIG. 16. Quantization errors (in dB) for Speaker CH for [ACC(1-15)], [M12(1-25)], and [ACC(1-15),M12(1-25)], as found for different codebook sizes.

sensors [ACC(1-15)] and [M12(1-25)], using the 10-bit codebook. As expected, the recognition accuracies were lower with the 10-bit codebook than with the 8-bit codebook: 83.9% vs. 86.3% for [ACC(1-15)] and 82.6% vs. 84.9% for [M12(1-25)]. We expect that increasing the amount of training data would alleviate this problem. For a limited database case such as ours, smoothing techniques may be used to improve the robustness of the HMM word models. However, because of the time and resource constraints of this project we were unable to investigate this possibility.

9.5.3.2 Tests of Simulated Single-Input Multisensor Systems

The long-vector approach for parallel-input multisensor systems has been demonstrated to lead to improved recognition performance. In this section, we report on some experiments that tested single-input multisensor configurations with our speech recognition research system. The research system allows more controlled experimentation than the Verbex system, which did not even train in some instances.

To compare the test performance of the single-input system (ACC, M12) with that of the parallel-input [ACC, M12] long vector, we performed several tests using three different simulated single-input (ACC, M12) systems in 105 dB noise for speaker RS. The results of these tests are shown in Table 29.

The first system tested was the same digital mix used to test (ACC, M12) with the Verbex 4000 (see Section 7.3.3). After noise was digitally added to RS's M12 data in 95 dB to achieve an effective 105 dB noise level, the M12 waveforms were highpass-filtered at 1500 Hz before the ACC and M12 waveforms were combined in the correct proportion; no highpass-filtering of the ACC signal was performed. We designate this system, tested using all 25 spectral parameters for the mix, as [(ACC, M12)(1-25)]. Referring to Table 29, we find that the performance of [(ACC, M12)(1-25)] is about the same as that of [M12(1-25)], slightly worse than that of [ACC(1-15)], and significantly poorer than that of the long-vector [ACC(1-15), M12(1-25)].

Next, we simulated the single-input system (ACC, M12) by using the long vector,

SENSOR CONFIGURATION	RECOGNITION ACCURACY
[ACC (1-15)] [M12 (1-25)] [(ACC, M12) (1-25)]	88.3% 87.2% 86.9%
[ACC (1-15), M12 (16-25)]	86.6%
[(ACC, M12)] , USING AVERAGE OF ACC AND M12 FOR BANDS 1-15 AND M12 FOR BANDS 16-25.	88.3%

Table 29. Tests of simulated single-input (ACC, M12) systems performed for speaker RS in simulated 105 dB ambient noise.

[ACC(1-15), M12(16-25)]. The choice of bands used for each sensor is akin to lowpass-filtering ACC and highpass-filtering M12 at roughly 1900 Hz before combining the two signals. Recall that for (ACC, M12), M12 was highpass-filtered, but ACC was not lowpass-filtered. However, since ACC's energy above 1900 Hz is small relative to M12's, we expected this simulation to have the same performance as [(ACC, M12)(1-25)], described above; this, in fact, proved to be the case. The resulting recognition accuracy for the 30-word test was almost 8% lower than that of [ACC(1-15), M12(1-25)]. This result indicates that removing the low-frequency information for M12 for the sake of enhancing subjective quality may, in fact, be harmful for machine recognition.

Because M12's low-frequency information appeared to be useful for an (ACC, M12) system, we conducted another test in an attempt to improve performance without the use of a parallel-input long-vector for the two sensors. In this simulation, we replaced each of the 15 spectral bands used for ACC in the previous test with the average (Spectral Band(ACC) + Spectral Band(M12))/2, and retained bands 16-25 for M12 in the parameter vectors to be tested. The recognition accuracy for this system was 88.3%, which was slightly better than for [ACC(1-15), M12(16-25)] but still significantly poorer than the long vector [ACC(1-15), M12(1-25)]. It appears, therefore, that although the low-frequency information for M12 is useful, combining its parameters in the low-frequency spectral bands with ACC's by simple averaging is not effective; to achieve an improvement in performance over the best constituent sensor, each sensor's low-frequency parameters must be included in the long vectors separately, as in [ACC(1-15), M12(1-25)].

9.5.4 Summary of Results from the Long-Vector Approach

We can summarize the results of our tests of the long-vector approach as follows. First, the parallel-input two-sensor system consisting of ACC and M12 achieved a recognition accuracy that was substantially better than the accuracy achieved by either sensor alone. Second, we found that when ACC is used alone, all 25 of its spectral-band parameters should be used; on the other hand, when its parameters are combined with M12's parameters in long

vectors, bands 16-25 of ACC can be omitted from the long vectors with no loss in recognition accuracy. Finally, the recognition performance of the parallel-input two-sensor system was substantially better than the performance of the single-input two-sensor system.

10. SUMMARY AND CONCLUSIONS

In this research, we investigated both single-input and parallel-input multisensor systems. To provide a rationale for developing these multisensor systems, we performed long-term and short-term spectral analyses and an articulation index study of previously measured data of one male and one female speaker. The results of our subsequent investigation of the two types of multisensor systems are summarized below, separately for each type.

10.1 Single-Input Multisensor Systems

We developed a spectral shaping method for improving the performance of a two-sensor configuration consisting of an accelerometer and a gradient microphone. Also, we developed several additional multisensor configurations, including a two-microphone system. A selected set of two-sensor systems and individual sensors were tested in 95 dB and 115 dB levels of simulated F-15 fighter aircraft cockpit noise, using formal speech intelligibility (DRT) and quality (10-point rating) tests. The test results show the spectral shaping method to be ineffective. The two-sensor systems tested produce essentially the same DRT scores and quality ratings in 95 dB and much higher DRT scores and quality ratings in 115 dB, as compared to the constituent individual microphones. Therefore, for high-noise applications involving human listeners, the two-sensor systems are clearly superior to any single microphone.

We tested and compared the performance of the various two-sensor systems and the individual sensors in speaker-dependent, isolated-word speech recognition. We used the commercial recognizer Verbex 4000 and three different vocabularies: a 20-word TI vocabulary (95 dB and 115 dB fighter aircraft cockpit noise), a 25-word minimal pairs vocabulary (95 dB noise), and a 13-word minimal pairs vocabulary (105 dB noise). In noise levels higher than 95 dB, the Verbex unit did not train and test successfully for many of the cases involving microphones. Of the two 95 dB cases, the 25-word minimal pairs vocabulary

was found to have a durational problem (see Section 7.3.5) that limited the achievable recognition performance. Because of these problems, we cannot make strong definitive statements comparing the performance of the two-sensor systems with the performance of the individual sensors. We can, however, make the following conclusions. Since the accelerometer is relatively insensitive to acoustic background noise, its recognition performance is essentially constant in different noise levels. For certain vocabularies (e.g., the 20-word TI vocabulary), the accelerometer also provides a good recognition accuracy (upper nineties for the TI vocabulary). For these cases, we suggest the use of a gradient microphone in low noise (say, below 100 dB) and the accelerometer in high noise, for achieving the best performance. For vocabularies involving discrimination only among unvoiced consonants, the accelerometer performs poorly in recognition as compared to gradient microphones. Even in these cases, the accelerometer would outperform the gradient microphone in sufficiently high noise levels. As a reasonable compromise, we suggest the use of a two-sensor system involving the accelerometer and a gradient microphone, with the provision that we filter the two-sensor signal using a lowpass filter with a 5 kHz cutoff in 95 dB noise and progressively lower cutoffs in higher levels of noise; in very high noise levels (e.g., 115 dB), the filtered signal may become almost the same as the accelerometer signal.

10.2 Parallel-Input Multisensor Systems

We demonstrated the feasibility of parallel-input multisensor speech recognition, using selected phonetic discrimination tests. We extracted features from individual sensor signals and determined the best overall case for each phonetic discrimination test by selecting the best of individual sensors and parallel-input two-sensor systems. In 105 dB noise, the feature-based recognition approach produced recognition accuracies, for different phonetic discriminations, in the range 88.1% - 94.9% for a gradient microphone and in the range 96.7% - 99.1% for the best overall case, with the latter case reducing the recognition errors to between one-half and one-twelfth of the number in the former case.

As a simple way of taking advantage of multiple, parallel inputs and to be able to use an

existing recognition algorithm, we then investigated a long-vector approach. In this approach, we formed, on a frame-by-frame basis, a long vector of parameters by merging the parameters of the parallel inputs provided by the individual sensors and evaluated the long-vector data using a discrete hidden Markov model-based speech recognition system. In our isolated-word recognition tests, we used a difficult 30-word minimal pairs vocabulary spoken by two talkers, a male in 105 dB simulated F-15 aircraft cockpit noise and a female in 95 dB noise. The results of our tests show that the parallel-input two-sensor system consisting of a throat accelerometer and a gradient microphone produced a recognition accuracy of 92.2% for 105 dB and 93.5% for 95 dB; the accuracies of the constituent sensors were, respectively, 85.0% and 86.6% for the gradient microphone and 87.8% and 85.8% for the throat accelerometer. Compared to the gradient microphone, the two-sensor system cut the recognition errors almost in half in both cases.

We have thus demonstrated the feasibility of parallel-input multisensor speech recognition and developed a simple and effective way of using multiple, parallel inputs with an existing recognition algorithm. Further research is warranted for modifying the recognition algorithm to exploit fully the parallel inputs and hence to achieve even more impressive gains in the recognition accuracy than the ones we reported above.

11. REFERENCES

1. V.R. Viswanathan, K.F. Karnofsky, K.N. Stevens, and M.N. Alakel, "Multisensor Speech Input", Final Technical Report RADC-TR-83-274, Contract No. F30602-82-C-0064, Bolt Beranek and Newman Inc., December 1983.

2. V.R. Viswanathan, K.F. Karnofsky, K.N. Stevens, M.N. Alakel, "Multisensor Speech Input for Enhanced Noise Immunity to Acoustic Noise", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Diego, CA, March 1984, pp. 18A.3.1-18A.3.4.

3. R.S. Nickerson, D.N. Kalikow, and K.N. Stevens, "Computer-Aided Speech Training for the Deaf", *J. Speech and Hearing Disorders*, Vol. 41, February, 1976, pp. 120-132.

4. V.R. Viswanathan and W.H. Russell, "New Objective Measures for the Evaluation of Pitch Extractors", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, FL, March 1985, pp. 411-414, Paper No. 11.10.

5. W.D. Voiers, A.D. Sharpley, and C.J. Hehmsoth, "Research on Diagnostic Evaluation of Speech Intelligibility", Tech. Report AFCR-72-0694, TRACOR Inc., January 1973.

6. R.M. Schwartz, "Acoustic-Phonetic Experiment Facility for the Study of Continuous Speech", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Philadelphia, PA, April 1976, pp. 1-4.

7. S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 4, August, 1980, pp. 357-366.

8. J. Makhoul, S. Roucos, and H. Gish, "Vector Quantization in Speech Coding", *Proc. IEEE*, Vol. 73, No. 11, November, 1985, pp. 1551-1588, Special Issue on Man-Machine Speech Communication.

9. L.E. Baum and J.A. Eagon, "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model of Ecology", *Amer. Math Soc. Bulletin*, Vol. 73, , 1967, pp. 360-362.

10. L. Rabiner, S.E. Levinson, and M.M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent Isolated Word Recognition", *Bell System Tech. J.*, Vol. 62, April, 1983, pp. 1075-1105.

Appendix A

DETAILS OF THE MULTISENSOR TEST DATABASE

Sentences

Print a draft of this letter.
I think someone's made a mistake.
Curt will put the box on my desk.
There's room for five more lines in the footnote.
Judy will lose the game of chess.
We need to make a bowl of soup.

20-Word TI Database

yes,	**no,**	**erase,**	**rubout,**	**repeat,**
go,	**enter,**	**help,**	**stop,**	**start,**
one,	**two,**	**three,**	**four,**	**five,**
six,	**seven,**	**eight,**	**nine,**	**zero.**

Minimal-Pair Words (ordered randomly)

met,	**psalm,**	**heed,**	**doze,**	**leaf,**
doak, (rhymes with soak)	**sod,**	**hid,**	**goad,**	**sin,**
leave,	**code,**	**bet,**	**debt,**	**sob,**
get,	**lease,**	**hood,**	**dode,** (rhymes with node)	**dope,**
let,	**sawn,** (rhymes with lawn)	**head,**	**mode,**	**shod,**
bed,	**ret,** (rhymes with let)		**net,**	**song,**
load,	**dote,**	**wet,**	**pode,** (rhymes with toad)	**dose,**
yet,	**fin,**	**had,**	**toad,**	**pet,**
hud, (rhymes with bud)	**hod,**	**node,**	**sog,** (rhymes with dog)	**bode.**

Appendix B

ACOUSTIC-PHONETIC FEATURES

The following frame-by-frame parameters played a part in the useful acoustic-phonetic features found in the APEF experiments:

1. CM75 -- The frequency (in Hz) above which 75% of the energy in the pre-emphasized speech spectrum lies.
2. F2M -- The second formant, after 3-point median smoothing.
3. F3M -- The third formant, after 3-point median smoothing.
4. LEZ -- The low-frequency energy, measured over a 120-440 Hz range, which has been smoothed with a 3-point zero-phase filter.
5. MEPZ -- The energy in the pre-emphasized LPC spectrum measured over a 640-2800 Hz range, which has been smoothed via a 3-point zero-phase filter.
6. R0P -- The energy in the pre-emphasized spectrum.
7. R1X -- The first autocorrelation coefficient, in dB.
8. ZC -- Number of zero crossings over a frame. The DC component is removed before ZC is calculated.

The acoustic-phonetic features we investigated are listed below. After the name of each feature, the number of the figure that should be referred to is given. If the definition of a feature is based on another feature in the list, the definition number of the second feature is given after its name. The words "initial" and "final" refer to cases where the discrimination was performed for initial and final consonants, respectively. Note that the figures which show energy contours from the Vought microphone are given merely to show the approximate regions, relative to the vowel in the utterance, where the features are measured. The energy contours in the figures do not necessarily correspond to the same words or parameters involved in a given feature.

1. burstdiff (Fig. 18) : LEZ at point F, where the plosive release would be expected - LEZ at point D, the time of LEZ's minimum value over the range from the vowel to point F.

   Used in NAS-VP/F test.

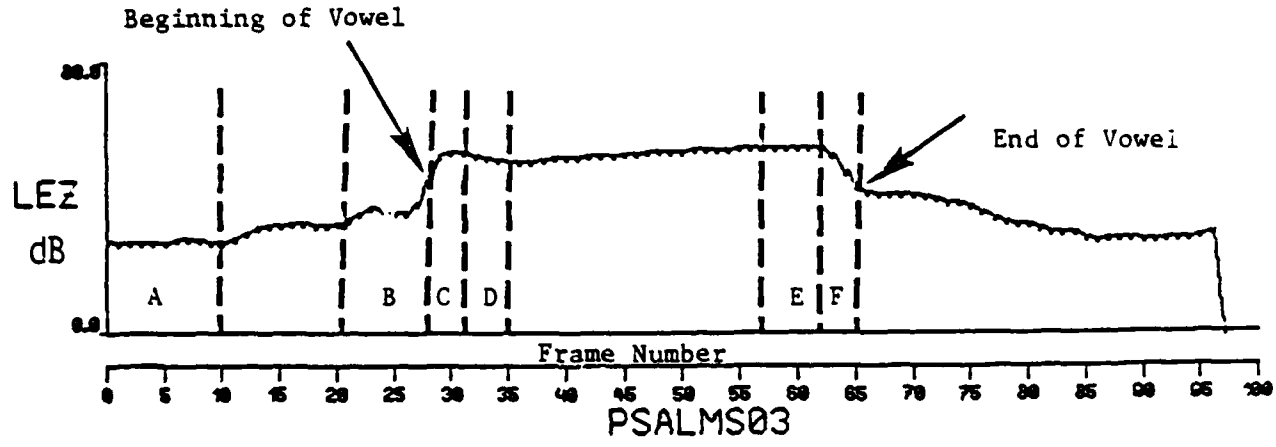


FIG. 17. Low-frequency energy contour for "PSALM" spoken by RS in 95 dB.

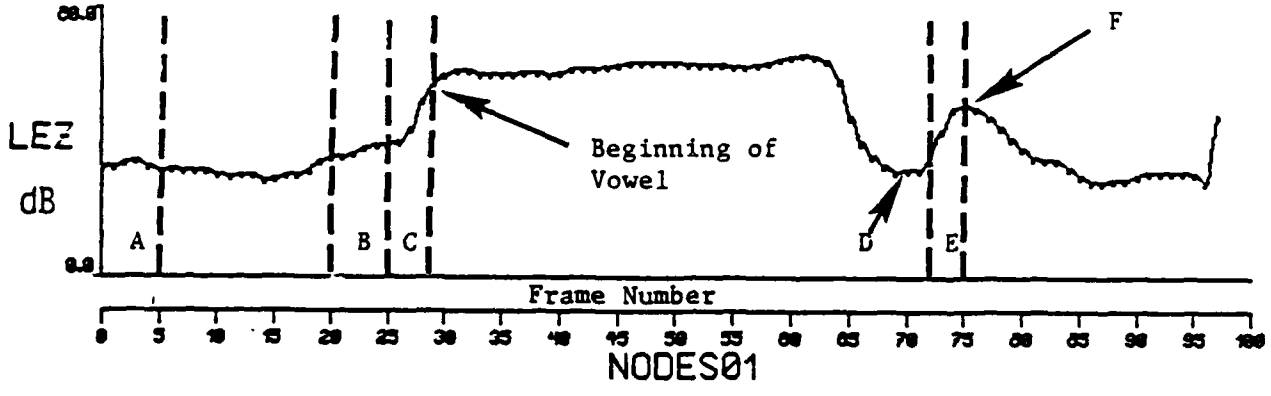


FIG. 18. Low-frequency energy contour for "NODE" spoken by RS in 95 db.

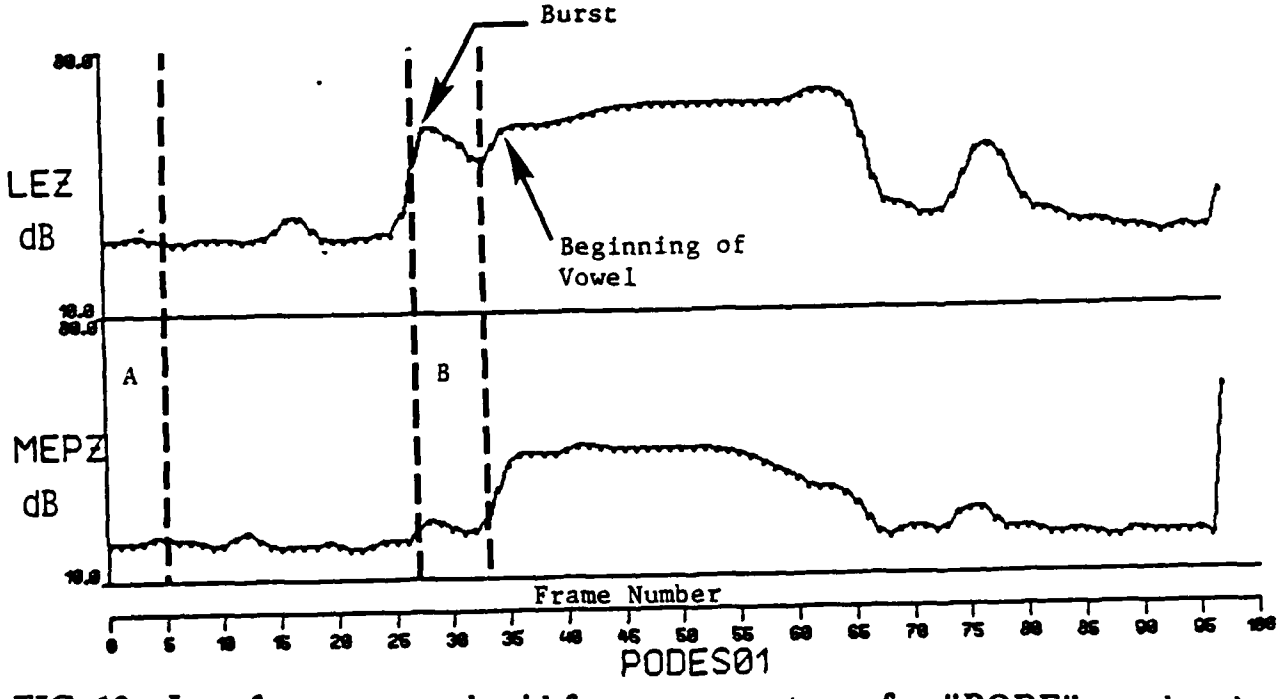


FIG. 19. Low-frequency and mid-frequency contours for "PODE" spoken by RS in 95 dB.

This value tended to be larger for voiced plosives than for nasals, because usually voiced plosives had an energy peak at F and a drop at D, while the nasals' energy stayed fairly level or dropped slightly from D to F.

2. cdiff (Fig. 17) : cvow(5) - ctrans, where ctrans = average CM75 in region C (initial) or F (final), during the CV or VC transition.

   Used in M-N/I, B-D-G/I, and B-D-G/F tests.

3. consilmepz (Fig. 18) : consmepz - silmepz, where consmepz = average MEPZ in region B, and silmepz = average MEPZ in region A.

   Used in B-D-G/I test.

   Consilmepz tended to be larger for [d] than for [b] and [g].

4. consent energy (Fig. 18) : average LEZ in region B (initial) or E (final)

   Used in NAS-VP/I and NAS-VP/F tests.

   In initial position, this value tended to be larger for nasals than for voiced plosives; in final position, the opposite was true.

5. cvow (Fig. 17) : average CM75 in region D (initial) or E (final)

   Used in B-D-G/F test.

6. diff (Fig. 18) : egy in çons - leznoise, where egy in cons = average LEZ in region B (initial) or E (final), and leznoise = average LEZ in region A.

   Used in NAS-VP/I and NAS-VP/F tests.

   In initial position, this value tended to be larger for nasals than for voiced plosives; in final position, the opposite was true.

7. f2diff (Fig. 17) : f2vow(9) - f2trans(8)

   Used in M-N/I, B-D-G/I, and B-D-G/F tests.

8. f2trans (Fig. 17) : average F2M in region C (initial) or F (final)

   Used in B-D-G/I and B-D-G/F tests.

9. f2vow (Fig. 17) : average F2M in region D (initial) or E (final)

   Used in M-N/I, B-D-G/I, and B-D-G/F tests.

10. f3diff (Fig. 17) : f3vow(12) - f3trans(11)

    Used in M-N/I, B-D-G/I, and B-D-G/F tests.

11. f3trans (Fig. 17) : average F3M in region C (initial) or F (final)

    Used in B-D-G/I and B-D-G/F tests.

12. f3vow (Fig. 17) : average F3M in region D (initial) or E (final)

    Used in M-N/I, B-D-G/I, and B-D-G/F tests.

13. ldiff (Fig. 18) : lstartavg - lsilavg, where
    lstartavg = average LEZ in region B, and
    lsilavg = average LEZ in region A.

    Used in VP-UVP/I test.

    Ldiff tended to be larger for unvoiced plosives than for voiced plosives, probably because of increased puff noise.

14. lezclose (Fig. 18) : average LEZ in region C

    Used in VP-UVP/I test.

    Because of pre-voicing of initial voiced plosives, lezclose for ACC tended to be larger for voiced plosives than for unvoiced plosives.

15. lezdiff (Fig. 19) : lezatb - leznoise, where
    lezatb = value of LEZ at burst, and
    leznoise = average LEZ in region A.

    Used in P-T-K/I test.

    Lezdiff tended to be largest for [p] and smallest for [t].

16. lvsmdiff (Fig. 18) : ldiff(13) - mdiff, where
    mdiff = mstartavg - msilavg, where
    mstartavg = average MEPZ in region B, and
    msilavg = average MEPZ in region A.

    Used in VP-UVP/I test.

    Lvsmdiff tended to be larger for unvoiced plosives than for voiced plosives, probably because of increased puff noise.

17. mepzdiff (Fig. 19) : mepzatb - mepznoise, where
    mepzatb = value of MEPZ at burst, and
    mepznoise = average MEPZ in region A.

Used in P-T-K/I test.

Mepzdiff tended to be largest for [t] and smallest for [p].

18. nasaccdiff (Fig. 18) : LEZ at point F - average LEZ in region A

    Used in NAS-VP/F test.

    Because nasalization was usually much stronger at point F for nasals than for voiced plosives, nasaccdiff, found with the nasal accelerometer, was usually larger for nasals than for voiced plosives.

19. slope (Fig. 18) : average derivative of LEZ in region C

    Used in NAS-VP/I test.

    This value tended to be smaller for nasals than for voiced plosives.

    r0pdiff (Fig. 19) : r0patb - r0pnoise, where
    r0patb = value of R0P at burst, and
    r0pnoise = average R0P in region A.

    Used in P-T-K/I test.

    This value tended to be largest for [t] and smallest for [k].

20. r1xdiff (Fig. 19) : r1xatb - r1xnoise, where
    r1xatb = value of R1X at burst, and
    r1xnoise = average R1X in region A

    Used in P-T-K/I test.

    This value tended to be largest for [p] and smallest for [t].

21. vot (Fig. 19) : vow - burst, where
    vow = time of maximum derivative of MEPZ over region B, and
    burst = time of burst.

    Used in P-T-K/I test.

    Vot tended to be largest for [p] and smallest for [t].

22. zcdiff (Fig. 19) : zcnoise - zcatb, where
    zcnoise = minimum value of ZC in region A, and
    zcatb = value of ZC at burst.

    Used in P-T-K/I test.

    Zcdiff tended to be largest for [p] and smallest for [t].

END

1-87

DTIC